The Retail Habitat

Toomas Laarits[*] Marco Sammon[†]

First version: May 9, 2022

This version: October 4, 2024

Abstract

Retail investors trade hard-to-value stocks. We document a large and persistent spread in the stock-level intensity of retail trading, even allowing for known biases in the attribution of retail trades. Stocks with a high share of retail-initiated trades are composed of more intangible capital, have longer duration cash-flows and a higher likelihood of being mispriced. Consistent with retail-heavy stocks being harder to value, we document that such stocks are less sensitive to earnings news, more sensitive to retail order flow and are particularly expensive to trade around earnings announcements. Overall, our findings document a new dimension of investor heterogeneity and suggest a comparative advantage of retail in trading hard-to-value stocks.

[*] New York University, Stern School of Business, tl1221@stern.nyu.edu.

[†] Harvard Business School, msammon@hbs.edu. We thank Yakov Amihud, Robert Battalio, Gregory Eaton (discussant), Alexi Savov, Eric So (discussant), Erik Stafford and Luis Viceira as well as the seminar audiences at Dartmouth, NYU Stern, the 2024 Midwest Finance Association and the 2024 NBER Behavioral Finance Working Group for comments and feedback on the paper. We thank Niels Gormsen, Eben Lazarus, Byungwook Kim and Andrei Golubov for sharing data and Myeongrok Doh for outstanding research assistance.

The cross-section of asset returns remains the principal venue for discriminating between theories of risk and return. With the well-established failure of the CAPM, a plethora of factors have been suggested to account for cross-sectional variation in returns instead. A closely related stream of research has studied the portfolios of different types of investors. Reflecting the failure of the CAPM, prior research has documented substantial heterogeneity in investor portfolios going well beyond differences in the share devoted to the market portfolio as prescribed by the CAPM. On a theoretical level, existing work has emphasized various investor characteristics such as risk aversion, elasticity of intertemporal substitution, investment horizon, financial sophistication, and background risk as important drivers of portfolio choice (Curcuru et al., 2010). More practically, an active empirical strand of work starting with Koijen and Yogo (2019) focuses on the identity of the investor—for instance banks, insurance companies, pension funds, investment advisors, and so on—and on each group's demand for assets with various characteristics.

In this paper we argue that the contrasting features of institutional and retail investors offer one particularly useful dimension to get a handle on such cross-sectional differences in trading and holdings. Institutional investors are frequently conceptualized in the finance literature as "smart money": they are typically much larger than any individual investor and therefore have the scale to acquire and process various kinds of data. Their scale also allows them to better overcome the fixed costs of trading, to take on leverage and short assets. Retail investors, by contrast, are seen as suffering from a litany of behavioral biases and cognitive errors, as less equipped to carry out meaningful research, and hence conceptualized as mere "noise traders" in the sense of Black (1986). All the while, retail investors have advantages of their own. As they are investing their own money, retail investors have more control over the investment horizon, and they do not face flow sensitivity to recent performance. They are also not constrained by mandates restricting the investable set or tying their compensation to performance relative to a specific benchmark.

Indeed, a number of empirical findings cut against a pure "noise trader" view of retail investors. A prominent strand of work finds that retail trades on the aggregate predict returns going forward with a positive sign.[1] On the other side of the retail-institutional divide, recent work in Di Maggio et al. (2019) documents that institutional investors trade away from stocks that are about to make earnings announcements, at times where their presumptive informational advantages ought to be

[1]See Kaniel et al. (2008), Barrot et al. (2016), Boehmer et al. (2021), among others.

the strongest. Further, the pandemic-era surge in retail trading (Welch, 2022) and the response of the stock market to stimulus checks (Greenwood et al., 2023) serve as reminders of the potentially large effect of retail traders on asset markets.

These findings, taken together, suggest that a conceptualization that goes beyond an informed *versus* uninformed dichotomy is required to account for the holdings and trading patterns of retail and institutional investors. Our argument is that the specific nature of the stocks heavily traded by retail and institutional investors plays a critical role. Specifically, we argue that the above empirical findings reflect the tendency of retail investors to trade stocks that are *hard-to-value*: stocks whose value is weakly tied to fundamentals, for instance, because they have long duration cashflows or have high shares of intangible capital. Within this hard-to-value set the presumptive informational advantages of institutional investors are mitigated. Indeed, we find that such hard-to-value stocks have lower sensitivities to fundamental news, higher sensitivities to past retail order flow, and see larger avoidance by institutional investors around earnings announcements. Taken together, our findings suggest a novel reason for the apparent large impact of retail investors on equity markets and spell out a distinct economic role for retail investors: holding risk in stocks where institutional investors informational advantage is mitigated.

To motivate this idea, we provide a model in the spirit of Kyle (1985). The model features an informed institutional investor with the capacity to research and generate information about one stock in the economy. Retail order flow is modelled as unpredictable and unrelated to fundamentals, but retail traders' behavior reflects an element of sophistication in that their trading intensity is allowed to depend on stock characteristics.

Where in the cross-section should this informed investor choose to produce information? At first blush it may seem that a stock with high retail trader presence would be the best bet, as the informed investor will find it easier to hide their trades among the retail investors' order flow. This logic only holds, however, if retail investors' propensity to trade is equal across stocks. Suppose, however, that retail investors tend to trade in hard-to-value stocks, defined as stocks for which the informed investors' research would yield the lowest amount of incremental information. Then, even conditional on a high level of noise-like trading activity, the informed investors may want to avoid learning about such stocks, as their expected profits would be lower on account of a weaker informational advantage. Overall, the model illustrates that where informed investors choose to trade depends on the relative strength of these two aspects: the intensity of retail trading, and the propensity of retail traders to focus on hard-to-value stocks.

Our empirical analysis is motivated by the tension between these two forces. First, to establish a necessary premise for our analysis—the existence of a subset of stocks with high retail trading interest—we document new facts on the distribution of retail trading in the cross-section, showing that retail trading intensity is both concentrated and persistent. In terms of magnitudes, almost 90% of stocks in the top 20% of retail trading intensity at any given point in time are still in the top two quintiles of retail trading intensity 12 months later.

Having documented a persistent retail focus, we seek to test which of the two forces suggested by the theoretical framework is more important empirically. We find that the concept of difficulty-to-value—defined in the model as the incremental fundamental signal precision gained from additional learning effort—is a particularly useful summary characteristic for explaining the cross-sectional heterogeneity in retail trading intensity: retail investors trade hard-to-value stocks.

To establish this cross-sectional regularity, we employ three types of proxies of difficult-to-value. Firstly, we examine cashflow duration (as constructed in Gormsen and Lazarus (2023)) under the view that firms with longer duration cashflows are harder to value because investors need to forecast fundamentals further in the future. Secondly, we examine various measures of intangible capital (Peters and Taylor, 2017; Kogan et al., 2017), which by nature is harder to value than physical capital (see e.g., Lev and Gu (2016)). Thirdly, we examine certain composite measures, the mispricing score of Stambaugh and Yuan (2017), the valuation uncertainty score of Golubov and Konstantinidi (2021), and the hard-to-value score of Ben-David et al. (2023). Across all three sets of measures we find that hard-to-value stocks see higher retail trading intensity: high retail stocks have longer duration cashflows, more intangible capital, more expected mispricing and more valuation uncertainty. Nearly all of these relationships hold both within the bottom and top 20% of firms by market capitalization, suggesting the relationship between retail investor activity and firm size is not driving our results. Combined with the evidence on retail trading persistence, these results establish a new way to capture the trading patterns of retail *versus* institutional investors.

The cross-sectional results described so far establish a novel regularity in the types of assets where retail trades make up a large share of total trading. In the second set of results, we demonstrate that this "retail sort" is particularly powerful in capturing differential return dynamics around earnings announcements in a way that is consistent with our contention that retail investors trade hard-to-value stocks. We find that high retail stocks have more volatile announcement news and returns, with a standard deviation of standardized unexpected earnings (SUE) that is more than three times as large for high retail stocks than low retail stocks. We also find that the dispersion in analysts' forecasts for high retail stocks is almost five times as large as for low retail stocks. This effect does not appear to be driven by analyst selection: analysts produce less accurate forecasts for high retail stocks, conditioning on a measure of an analyst's attention constraints and controlling for heterogeneity in analyst skill as well as firm characteristics.

These results regarding earnings announcements are in line with our claim that retail traders concentrate in stocks where the link between observable fundamentals and market prices is relatively weaker. To quantify this connection, we employ earnings-response regressions following Kothari and Sloan (1992) and find that, for a given magnitude earnings surprise, high retail stocks' prices respond significantly less to earnings news than low retail stocks. A stock in the highest quintile in terms of past retail trading share has a roughly 40% lower sensitivity to standardized unexpected earnings news than a stock in the middle quintile. This effect is unchanged by controlling for a litany of characteristics known to be correlated with retail activity and holds at almost every point along the firm size distribution.

The two sets of results described so far establish descriptive differences in the composition of high retail portfolio, as well as dynamics around earnings announcements. In the third set of results, we provide evidence that these differences across the retail sort reflect proactive decisions on behalf of retail investors as a group. Specifically, we document that retail-heavy stocks see substantial retail inflows in anticipation of earnings announcements. Cumulative net buying by retail investors (normalized by respective daily trading volume) adds up to over 2% in the immediate run-up to earnings announcements. Retail traders hence hold a disproportionate share of the earnings news risk of such stocks. Our result elucidates the phenomenon documented in Di Maggio et al. (2021) regarding the tendency of institutional investors to exit positions ahead of earnings announcements: that this effect is substantially larger among high retail stocks suggests institutional investors understand that hard-to-value stocks have volatile and idiosyncratic earnings-day returns and want to avoid exposure to such risk.

The average net trading of retail investors around earnings announcements is strongly suggestive of liquidity provision: retail investors as a group trade into announcing stocks in the last week before a scheduled announcement—at exactly the time that institutions tend to exit. The announcement-by-announcement direction of retail order flow confirms a substantial role for liquidity provision. On the announcement level, we find strong evidence that stocks heavily bought by retail investors in anticipation of earnings news releases outperform stocks heavily sold by retail investors over the subsequent 60 days, inclusive of the earnings day itself. This pattern, documented unconditionally by Kaniel et al. (2012), is particularly pronounced for stocks with high past levels of retail trading. In a decomposition exercise, we attribute about half of this predictable return differential to liquidity provision, emphasizing the active role retail traders take around earnings announcements in this subset of stocks.

Despite the liquidity provided by retail investors, the aggregate trading conditions for high retail stocks deteriorate around earnings announcements. We find that high retail stocks have abnormally high bid-ask spreads in a 5-day window around the scheduled event, relative to the stock-level average over the past month. In terms of magnitudes, the abnormal effective spread of Holden and Jacobsen (2014) is 4 basis points higher on the earnings announcement day itself for high retail stocks relative to the average stock. For reference, this increase is $2/3^{rds}$ the size of the average value-weighted bid-ask spread in 2021 of 6 basis points (Greenwood and Sammon, 2022). We interpret this finding through the lens of our model: liquidity in high retail stocks deteriorates around earnings announcements despite the higher share of retail trading. Our preferred interpretation is that—given our results on past retail trading's predictive power for future returns—retail investors' order flow is informative about future returns, and market makers increase trading costs to offset the risk of adverse selection. This can occur even if retail investors have no information about true fundamentals, but rather retail order flow is informative about future retail order flow in the same direction – an example of noise traders "making their own space" as described in De Long et al. (1990). Under this view, arbitrageurs may avoid leaning against retail orders, as they face the risk of future retail trades pushing prices further against their position.

In sum, our results on earnings announcements document two novel phenomena across the retail sort. One, the announcement time returns of high retail stocks are more volatile and exhibit a weaker relationship with earnings news. Two, retail investors take an outsized long position in these stocks before earnings announcements, consistent with liquidity provision. In the final set of results, we show that these two facts lead to substantial differences in average returns earned over the announcement window.

As established in a long literature starting with Beaver (1968), stocks tend to earn high average returns when they are scheduled to make earnings announcements. A potential explanation for the earnings announcer premium is that announcing firms provide information about non-announcing firms and therefore the premium is compensation for exposure to systematic risk, as argued in Savor and Wilson (2016). Our expectation, therefore, is that such a premium should be smaller for high retail stocks, as we have shown their earnings announcements to be mostly comprised of idiosyncratic information. This is precisely what we find: across the size distribution, high retail stocks consistently see lower announcement time returns. Unconditionally, stocks in the top quintile of market capitalization earn an earnings announcement premium of 18bps. However, among this set of large stocks, those in the highest retail trading quintile see an average return of negative 18bps over the same time window. In pre-announcement returns, the pattern is flipped: high retail earns higher returns across the size distribution. The aggregate version of this pattern reflects a known puzzle of high pre-announcement returns (Frazzini and Lamont, 2007), but again our findings establish a new cross-sectional regularity: this pre-announcement return is reliably stronger among the high retail stocks, across the size distribution. In other words, it's strongest for the stocks which institutional investors tend to trade away from in anticipation of announcements, suggesting liquidity provision as a potential driver of the return differential.

1.1 Attributing Retail Trades

Our proxy for retail trading intensity is based on the algorithm developed in Boehmer et al. (2021). Their work had immediate impact: as of June 2024 it has garnered 486 Google Scholar cites and has been employed in a variety of empirical studies (Battalio et al. (2023) lists 28 published papers in top accounting and finance journals). In parallel, a number of studies have raised concerns about mis-attribution of trades, with the algorithm both missing true retail trades and classifying institutional trades as retail. Barber et al. (2023) trade on their own account and carry out a forensic analysis of the resulting records in the TAQ database. They find that only 35% of their trades are classified as retail and 72% are signed correctly. In related work, Battalio et al. (2023) measure trades identified by the BJZZ algorithm against a proprietary dataset of actual retail orders. They find that only 28% of their sample of retail trades are classified by the BJZZ algorithm as retail, though they find better performance—94% success rate—in terms of assigning trades as buyer- or seller-initiated. Additionally, these papers document systematic patterns in the mis-attribution rates: both find that the success of the BJZZ algorithm is related to stock-level characteristics such as bid-ask spreads and nominal prices.

We seek to alleviate concerns about potential measurement error in a variety of ways. Firstly, we note that we are primarily interested in the ordinal ranking of stocks on retail interest, and that the directional attributes of misclassification identified by Barber et al. (2023) and Battalio et al. (2023) either bias down the retail share of trading precisely for the stocks that see a high level in our calculation, or else are quantitatively not large enough to overcome the spread we observe in retail share of trading volume. Secondly, we confirm that all our directional results hold essentially unchanged using the improved trade direction classification algorithm advocated by Barber et al. (2023) that uses the Lee and Ready (1991) midpoint method to assign buyer- or seller-initiated trades. Thirdly, we show that the retail trading intensity measure constructed following Boehmer et al. (2021) data is strongly correlated with lower-frequency proxies of retail trading and retail holdings. In all, our reading of the data is that with respect to our results, there is at most a small amount of directional bias introduced by the BJZZ algorithm, despite substantial noise in both trade identification and trade direction classification.

1.2 Connection to the Literature

Our work contributes to an active strand of research that has highlighted the importance of investor heterogeneity and less-than-perfect risk-sharing in determining the risk-return trade-off in security prices. One part of this work seeks to estimate demand curves of different investor classes as functions of various characteristics (Koijen and Yogo, 2019; Koijen et al., 2020; McLean et al., 2020; Haddad et al., 2021; van der Beck, 2022). Our work documents a new point of distinction in the trading habits of two principal investor classes: retail and institutional investors. Other recent work in Balasubramaniam et al. (2023) and Gabaix et al. (2022) has studied the portfolios of retail investors specifically. Balasubramaniam et al. (2023) use account-level data from India to document the role of characteristics in attracting retail holdings. They find that firm age and nominal price, and, to a weaker degree, turnover and recent returns are the characteristics that best capture the heterogeneity in retail holding intensity. Our aggregate retail trading data is consistent with a retail focus on firm age and nominal price, as well as turnover and past returns, while pointing to a unifying strand underlying these regularities.

Outside of that recent work, the literature on retail investors has devoted surprisingly little attention to the determinants of retail trading and holdings in the cross-section. Most of the existing literature has focused on various behavioral frictions that bring stocks to the attention of retail investors. However, we find that there is substantial and persistent cross-sectional heterogeneity in retail trading intensity, and it can be explained by a metric which is not obvious from looking at past returns, betas, or accounting figures alone. Our results add to the literature by suggesting that difficult-to-value stocks attract particular retail attention or, equivalently, deter institutional investors.

Indeed, this aspect of retail selection allows us to reconcile two broad, seemingly contradictory aspects of retail investing. On one hand, research has repeatedly found that retail trades – on aggregate – tend to positively predict stock returns going forward. For example, Kaniel et al. (2012) show that the direction and magnitude of retail order flow predict returns on and after earnings announcements. Along the same lines, in more recent work, Welch (2022) documents that Robinhood investors as a group did well in 2020-21.[2] On the other hand, retail traders have been shown to suffer from a litany of behavioral biases including: excessive trading (Barber and Odean, 2000, 2002), familiarity bias (Huberman, 2001; Seasholes and Zhu, 2010), extrapolation (Benartzi, 2001) and the disposition effect (Odean, 1998; Dhar and Zhu, 2006; Vaarmets et al., 2019), to name a few. Moreover, relaxation of retail investors' budget constraints sees the prices of retail-heavy stocks rally (Greenwood et al., 2023). Taken as a whole, our results suggest that because of the selection by retail traders into hard-to-value stocks, these biases and predictable errors are particularly hard for professional investors to correct.

More broadly, our results can be used to recast several existing results in the asset pricing literature by emphasizing how the relative importance of two types of investors can directly contribute to these phenomena. First, previous literature has shown significant effects of retail investor buying on stock prices (Kumar and Lee, 2006; Greenwood et al., 2023). Our results on the concentration

[2]The evidence regarding retail investors' trading performance in options is mixed, with Bryzgalova et al. (2023) finding that retail investors lose money on average, while Bogousslavsky and Muravyev (2024) argue that average losses are small, and investors may use options as a relatively inexpensive way to access leverage.

of retail investor trading, as well as the types of stocks preferred by retail, may explain why retail investors can have such a large effect on prices despite their relatively small share of overall stock market wealth. Second, the focus on hard-to-value stocks can explain why retail order flow is a strong predictor of returns going forward, as documented in Kaniel et al. (2012). In fact, we show that such predictability is particularly pronounced within the set of high retail share stocks. Given that retail order flow is persistent, and that retail investors focus on stocks which are relatively more expensive to trade, it may be difficult for institutional investors to maintain bets against retail order flow long enough to benefit from long-run reversion (De Long et al., 1990). Further, we believe our results speak to the literature on how inventory risk borne by intermediaries can lead to high pre-earnings announcement returns (Johnson and So, 2018). Specifically, we find that high retail stocks have earnings-announcement returns which are significantly more volatile than low retail stocks. Market makers holding inventory therefore, may offer a premium to retail investors buying ahead of the announcement, as this would reduce their naturally long exposure to this risk from holding inventory. Finally, we show that the stocks which retail investors tend to favor have very low or high mispricing scores (constructed using the data from Stambaugh and Yuan (2017)), suggesting they are often in the extreme ends of anomaly portfolios. This opens the door for retail investors to directly contribute to anomaly returns, as retail investors' trading in these stocks makes it tougher for institutional investors to try to correct any mispricing.

2 Hypothesis development

In this section we outline a model and three sets of predictions that guide our empirical exercises.

2.1 Motivation

Consider a model in the spirit of Kyle (1985) with multiple securities and two periods. There are gains to specialization, and an informed insider, representing institutional investors, can generate a signal about the value of one stock. The securities themselves are heterogeneous in two ways: (A) the level of noise trading intensity, standing in for differences in the intensity of retail trading, and (B) the precision of the signal the insider can generate, standing in for the difficulty of valuing the stock. Within this model we ask: where in the cross-section would the institutional investor find it most profitable to produce information?

Reflecting standard intuition, all else equal, the insider's profit will be larger in stocks with higher noise trading intensity. One might expect, then, that institutional investors expend most of their attention learning about stocks with more retail trading activity. What this line of argument misses, however, is that retail trading activity need not be uniformly distributed in the cross-section of stocks. Instead, retail trading may be most concentrated in stocks where insiders have the worst quality of information, which in the model are the stocks for which the insider can generate a relatively less precise signal. If retail investors focus their trading in such hard-to-value stocks, then high retail stocks might offer worse overall expected profits to the insiders, as they face poor enough signal precision to outweigh the expected benefits of hiding their trade among retail order flow.

The two-period Kyle (1985) model described in Appendix A.1 allows us to make this point explicitly. We simulate the model and plot the insider's profit as a function of signal precision and noise trading intensity. Appendix Figure A1 shows that, unsurprisingly, the insider's profit is monotonically increasing in both signal precision and noise trading activity. The more surprising result is that the insider's profit can be lower in a high noise trading intensity stock than a low noise trading intensity stock, if the precision of their signal is sufficiently higher in the low noise trading intensity stock. In other words, which force dominates depends on whether or not retail investors have a persistent habitat of hard-to-value stocks.

In our baseline version of the model, retail investors' order flow is uncorrelated with securities' terminal payoffs. The retail investors exhibit some sophistication, though, in that they can pick particular stocks in the cross section. The results from the calibrated two-period model therefore suggest that which of these forces dominates—hiding among noise traders vs. precision of signals—is an empirical question.

2.2 Cross-sectional heterogeneity in retail trading intensity

Motivated by the model, we first seek to establish which of these two forces—hiding among retail order flow *versus* precision of signal—dominates. To this end, we employ a number of proxies for difficult-to-value and summarize them across the retail sort. As mentioned in the introduction, we document differences across cash-flow duration (Gormsen and Lazarus, 2023), intangible capital

(Peters and Taylor, 2017; Kogan et al., 2017), presence in mispricing portfolios (Stambaugh and Yuan, 2017), valuation uncertainty (Golubov and Konstantinidi, 2021) or other composite measures of difficulty to value based on trading volume, analyst dispersion, age and idiosyncratic volatility (Kumar, 2009; Ben-David et al., 2023). Consistently across proxies, we find that stocks with a higher share of retail trading are harder to value. We establish this result in unconditional sorts, as well as double sorting on size and the difficult-to-value proxy, and controlling for a litany of other characteristics that have been shown to correlate with retail interest: nominal price, past returns, and return volatility, among others. In addition, we discuss the importance of biases in the BJZZ algorithm documented by Barber et al. (2023) and Battalio et al. (2023) and show that the biases either work against the direction of the spread we find, or else are quantitatively minor compared to the observed spread in retail trading intensity.

Having shown that difficult-to-value is a powerful way to summarize retail trading interest, we develop three sets of predictions regarding the stocks heavily traded by retail investors.

2.3 Predictions on retail trading and earnings announcement dynamics

Our first set of predictions concerns the differences in earnings announcement dynamics, both regarding the earnings news as well as the resulting returns. Given the difficulty of producing a signal regarding the value of stocks heavily traded by retail, we anticipate such stocks to have larger earnings-day return volatility. As these firms are comprised of more intangible capital, we additionally anticipate their earnings to be harder to forecast, leading to both a larger magnitude of earnings news and to higher dispersion in analysts' earning forecasts. In sum, we predict that:

Prediction 1A: *High retail stocks should have more volatile earnings-day returns and earnings news. In addition, high retail stocks should have more dispersion in analysts forecasts.*

A potential concern with testing Prediction 1A is that there is selection in terms of which types of analysts cover high retail stocks and low retail stocks. If, for example, low quality analysts cover high retail stocks, such stocks may have larger earnings surprises even though they are just as hard to value as low retail stocks. However, we develop a test of prediction 1A which compares accuracy within the stocks a given analyst covers, still finding evidence of larger forecast errors among high retail stocks.

Given that high retail stocks are hard to value, any news about current cash-flows will have a relatively smaller effect on current prices. The logic is that for firms with long duration cash-flows, or a significant amount of their value in intangible capital, current earnings are not as relevant for present value. Additionally, in hard to value stocks, different investors may focus on different pieces of the news, leading to more disagreement (Hong and Stein, 2007) or investors may choose to ignore public signals altogether (Banerjee et al., 2021; Hirshleifer et al., 2009; Engelberg, 2008; Cohen et al., 2020), ultimately leading to under-reaction to news. These considerations yield the following prediction:

Prediction 1B: *High retail stocks should respond relatively less to earnings news. Earnings surprises for such stocks should be mostly driven by idiosyncratic news.*

2.4 Prediction on retail trading around earnings announcements

The first set of predictions concerned the dynamics of earnings news and returns around scheduled announcements. Our second prediction ascribes an active role to retail investors around these information events, particularly with respect to liquidity provision. Like with the first set of predictions, we consider the cross-section of stocks as a function of their prior intensity of retail-initiated trades. If, according to Prediction 1B, high retail stocks are less sensitive to earnings news, we anticipate the presumptive advantage of institutional investors in trading based on fundamental signals to be mitigated around these events. Additionally, prior work has documented that institutional investors as a group tend to trade away from announcing stocks (Di Maggio et al., 2021).

The combination of these two aspects of earnings announcements suggests that for high retail stocks, institutional investors should find their informational advantage to be weakest and reduce their trading intensity, as well as trade away from the announcing firms. Retail, in turn, should make up a larger share of trading volume in these stocks and hold a disproportionate position through earnings announcements. In other words, we hypothesize that the unconditional pattern documented in Di Maggio et al. (2021) regarding institutional investors trading away from announcing firms is due to them seeking to avoid unpredictable, idiosyncratic return exposures which are strongest in high retail stocks. In sum, we have the following testable hypothesis:

Prediction 2: *High retail stocks should have a particularly high share of retail trading intensity around earnings announcements. Around the scheduled earnings announcement, retail act as liquidity providers and hold an outsize position in announcing stocks.*

2.5 Prediction on retail trading and the earnings announcement premium

Finally, the differing information dynamics across the retail sort during the earnings cycle should be evident in average returns. Prior work in Savor and Wilson (2016) has argued that the announcement risk premium—the positive average returns earned by announcing firms—derives from the systematic value-relevant information in announcements regarding non-announcing firms. This mechanism is unlikely to apply to high retail firms: because these stocks are hard to value, the information contained in a given earnings announcement is likely to be mostly idiosyncratic and carry little information about the valuation of other stocks, as hypothesized in Prediction 1B. For that reason, we anticipate the earnings announcement premium to be decreasing in the intensity of retail trading:

Prediction 3: *High retail stocks should have a lower earnings announcement premium.*

3 Data

3.1 Measuring Retail-initiated Trades

In this section, we briefly describe our main data sources and variable construction. Our key measure of retail trading activity is $\text{RSVOL}_{i,t}$, the retail share of trading volume, defined as

$$\text{RSVOL}_{i,t} = \frac{\text{RBuy}_{i,t} + \text{RSell}_{i,t}}{\text{Volume}_{i,t}}, \tag{1}$$

where $\text{RBuy}_{i,t}$ and $\text{RSell}_{i,t}$ are the number of shares in retail-initiated buy and sell trades, respectively. $\text{Volume}_{i,t}$ is total daily volume on the TAQ tape. In words, $\text{RSVOL}_{i,t}$ is the fraction of stock i's total trading volume on day t accounted for by retail-initiated buys and sells. We report $\text{RSVOL}_{i,t}$ in percentage terms. In addition to a daily measure of retail-initiated trading, we also construct a monthly counterpart. For each month τ, we sum up the retail-initiated trades $\text{Rbuy}_{i,t}$ and $\text{Rsell}_{i,t}$ as well as total volume $\text{Volume}_{i,t}$ and then construct monthly $\text{RSVOL}_{i,\tau}$ according to Equation 1.

RSVOL is our preferred measure of retail trading intensity because it does not depend on the level of trading in a given stock. There is scant agreement on the economic forces underlying both the level and cross-sectional variation in trading volume but the construction of RSVOL sidesteps that issue.[3] That said, in the Online Appendix we confirm the robustness of the upcoming main empirical findings by sorting over retail-initiated turnover, by sorting over the reciprocal of institutional ownership share, and by double sorting first on turnover, then on retail trading intensity. We refer to these robustness checks throughout the paper in conjunction with the baseline results.

Retail trades are identified using the algorithm proposed in Boehmer et al. (2021) that relies on the regulation of U.S. security markets requiring disclosure on price improvement for retail-initiated trades that are internalized. We construct this measure using the TAQ millisecond data from 2007-2021.[4]

Our stock sample consists of all CRSP ordinary common shares that are traded on major exchanges and can be matched to the retail activity data. Specifically, we restrict to share codes 10-11 and exchange codes 1-3. For the mapping between TAQ and CRSP identifiers, we use the linking table provided by Wharton Research Data Services (WRDS).

To quantify cross-sectional differences in retail activity, each month, we sort securities into five groups based on retail trading intensity the prior month i.e., $\mathrm{RSVOL}_{i,\tau-1}$.Panel A of Figure 1 plots the time series of average $\mathrm{RSVOL}_{i,\tau}$ in the 1st and 5th quintiles of portfolios sorted on prior month $\mathrm{RSVOL}_{i,\tau-1}$. This figure shows that there is substantial cross-sectional heterogeneity in retail activity: the share of retail-initiated trades ranges from 2% to 20% going from the bottom to top quintile. As reported in the first column of Table 1, the average gap in retail trading intensity going from the top to bottom quintile is about 13 percentage points.

3.2 Quantifying the Bias in Identifying Retail Trades

In the short time since the publication of the BJZZ algorithm, a number of papers have argued that the method is prone to both false negatives—missing true retail trades—as well as false positives—classifying institutional trades as retail. In particular, Barber et al. (2023) trade on their own account and analyze the resulting entries in the TAQ database, finding that only 35% of their trades are classified as retail, and only 72% are signed correctly. Battalio et al. (2023) measure the BJZZ retail trade attribution against a proprietary dataset of true retail orders, and find that only

[3] See Liu et al. (2020) for a summary of the literature on the excess trading volume puzzle.

[4] Boehmer et al. (2021) note that from 1/2016-9/2018, the SEC's tick size pilot program likely affected the prevalence of subpenny price improvements among the included stocks.

28% of true retail orders are classified as retail and 94% are signed correctly.

We seek to alleviate concerns about measurement issues in three ways. Firstly, we note that in most of our applications we are interested in the ranking of stocks based on retail trading activity, and not in the true level of such trades. Any noise in the algorithm need not pose a hurdle for this application, insofar as such measurement error is sufficiently uncorrelated with underlying economic drivers of retail trading. The median stock at the end of 2021 saw over 350,000 monthly trades, while the median stock in the largest size quintile saw over a million trades per month. Even with an attribution rate of as low as 30% (as measured byBarber et al. (2023) and Battalio et al. (2023)), the measured share of retail-initiated trades would be highly correlated with the true share.

Employing the cross-sectional estimates of false negatives from Barber et al. (2023) and false positives and false negatives from Battalio et al. (2023) we report in Appendix A.2 that the documented biases on the attribution of retail trades either go against the direction of the spread we document, or are quantitatively small relative to the 13 percentage point gap in retail trading share going from the bottom to top quintile.

Secondly, in the results where we do use directional measures of retail trading, we show that our findings are essentially unchanged using the improved trade-signing algorithm advocated by Barber et al. (2023) that uses the Lee and Ready (1991) midpoint method. That the directional results are nearly identical despite the substantial share of misattribution documented in Barber et al. (2023) again suggests that the errors in the algorithm are mostly idiosyncratic and do not bias our estimates in any particular direction.

Thirdly, we document that retail trading intensity constructed using the BJZZ algorithm is strongly correlated with other, less frequently observed but more precise proxies of retail trading intensity and ownership. In Table 2, we show that our rankings based on retail trading intensity are strongly inversely correlated with institutional ownership from Form 13-F data. Yet another way of quantifying retail trading activity at a monthly frequency is by using SEC rule 605 reports filed by wholesalers, and in Appendix A.2 we show that our rankings are similar to those based on this regulatory data. In Appendix A.2 we also show our rankings are similar to those based on the number of Robinhood users in Robintrack data.

Overall, our reading of the combined evidence is that while the BJZZ algorithm suffers from measurement error, the misattribution of trades is mostly unsystematic, giving us confidence that the resulting ordinal ranking of stocks with respect to retail trading intensity is unbiased.

3.3 Earnings Announcement Data

A number of our predictions concern return dynamics around earnings announcements. To test these predictions we need to establish the first time investors could have traded on earnings information. We identify days when investor could have first traded on the earnings information during normal trading hours using the earnings release date and time in IBES. If earnings are released before 4:00 PM Eastern Time on a trading day between Monday and Friday, that day will be labeled as the effective earnings date. If earnings are released on or after 4:00 PM Eastern time between Monday and Friday, over the weekend, or on a trading holiday, the next trading date in CRSP is labeled as the effective earnings date. To be conservative, we instead use the first trading day on or after the release date of quarterly earnings (RDQ) in Compustat if it occurs at least one day before the date identified using IBES (Livnat and Mendenhall, 2006). We use the mapping file from WRDS to link IBES data to CRSP.[5] Because of our focus on earnings announcements we restrict the sample to firms for which we are able to construct earnings expectations.

3.4 Other Data

We additionally employ a bevy of proxies for difficult-to-value, as described in the introduction. For a detailed description of all the variables used in our analyses, see Section A.3 of the Appendix.

4 Retail Trading and Stock Characteristics

In this section we document significant cross-sectional dispersion as well as persistence in retail trading activity. We then examine stock-level characteristics that account for this heterogeneity, finding that stocks favored by retail traders can be characterized as relatively hard to value.

[5] At the start of our sample in 2007, IBES covers 88% of ordinary common shares traded on major exchanges in CRSP. This number declined slightly over time to 84% by 2020. The firms not covered by IBES tend to be smaller and younger on average.

4.1 Retail Trading Intensity in the Cross-Section

There is substantial heterogeneity in the intensity of retail-initiated trading in the cross-section of stocks. In the 2007 to 2021 sample, marketable retail orders identified by the Boehmer et al. (2021) algorithm make up 7.94% of daily total trading volume for the average stock. Our first set of results document that the cross-sectional variability of retail-initiated share of volume, denoted $\text{RSVOL}_{i,t}$, is large relative to its unconditional mean. To establish this spread in retail trading intensity, each month, we sort securities into quintiles based on retail trading intensity in the prior month i.e., $\text{RSVOL}_{i,\tau-1}$. In order to quantify differences across the retail sort, we estimate regressions of the form:

$$\text{Outcome}_{i,\tau} = a + \beta_1 1_{i\in Q1_{\tau-1}} + \beta_2 1_{i\in Q2_{\tau-1}} + \beta_4 1_{i\in Q4_{\tau-1}} + \beta_5 1_{i\in Q5_{\tau-1}} + \epsilon_{i,\tau} \tag{2}$$

where $1_{i\in Qj_{\tau-1}}$ are indicator variables for whether stock i was in retail trading intensity quintile j in the previous month $\tau - 1$. The omitted group is the middle quintile of retail trading intensity. Standard errors are double clustered at the stock and month levels and we include year-month fixed effects.

In the first three columns of Table 1 we show the moments of $\text{RSVOL}_{i,t}$, the retail-initated share of trading volume. The gap in retail trading share between high and low retail stocks is about 13%. The second and third columns restrict the sample to the smallest and largest quintiles in terms of market capitalization, respectively. The gap in retail trading intensity is present for both small and large stocks, with respective sizes of 15% and 11%, as calculated in the table footer.

The other two sets of three columns repeat this analysis for total share turnover and retail-initiated share turnover. Both turnover and retail-initiated turnover are measured as the number of shares traded, normalized by shares outstanding and reported in percentage terms. The gap in share turnover going from low to high retail is about 8%, while the gap in retail-initiated turnover is about 3%. Both of these measures see larger gaps across the retail sort when restricting the sample to small stocks, though the differences across high and low retail stocks are statistically significant in all specifications: we report a formal test of of equality between the coefficients on Q1 and Q5 in the table footer. Again, the differences in turnover and retail initiated turnover are present for both small and large stocks.

The finding that high retail share stocks also have high turnover raises the possibility that sorting on retail share of trading volume is just another way of sorting on overall turnover. In Appendix A.4 we perform a double sort on overall turnover and retail-initiated turnover. We show that within each portfolio formed on turnover, the sub-portfolios formed on retail-initiated turnover have similar retail shares of trading volume as in the baseline unconditional sort. In other words, the variation in the retail share of trading volume is coming from retail trading itself, not a lack of trades by institutional investors.

The time-series dimension of average retail share is illustrated in Panel A of Figure 1. Here we plot the equal-weighted average retail intensity within the top and bottom quintiles of past retail intensity. For high retail stocks (Q5), retail investors have become an an increasingly large fraction of trading volume, now at around 20% of total shares traded. For low retail stocks (Q1) retail intensity has been relatively stable at about 2% of total trading, indicating the the overall increase in retail trading intensity has been focused in the high retail stocks, the focus of our analysis.

Finally, the retail sort is persistent over time. In Appendix A.5, Table A3 shows the 12 month transition probabilities across RSVOL-sorted bins. As Panel A shows, stocks in the highest quintile in terms of retail share of trading have a 66% probability of remaining in the top quintile 12 months in the future. These same stocks have an almost 90% probability of remaining in one of the top two retail-heavy portfolios.

In light of the striking persistence, what could lead stocks to transition from low to high retail? Our main argument provides one potential answer: firms might become harder to value, sometimes abruptly. Two recent examples illustrate this point. Hertz saw close to 30% of its trading volume stem from retail investors when it filed for bankruptcy in 2020. Similarly, First Republic Bank was a low retail interest stock until the regional banking crisis of 2023. Our reading of these two examples is that they launched stocks that had relatively regular, predictable businesses into hard-to-value territory.

4.2 Stock Characteristics across Retail Portfolios

The results in the previous section establish substantial heterogeneity as well as a strong degree of persistence in retail trading intensity. The combination of these features opens up the possibility of a retail habitat—the focus of retail in a particular set of assets. Our first main result is establishing that this habitat can be summarized as hard-to-value.

We document standard firm characteristics across RSVOL quintiles in Table 2. We find that high retail stocks are smaller, younger, have low nominal prices, low recent returns (measured from month -12 to month -2), higher book-to-market ratios and tend to have low or negative earnings yields. Note that the first column reports a median regression to document that the typical firm in the high retail bucket is a small firm. There are a number of very large firms, though, in the high retail quintile (e.g., Tesla) and for that reason the average (mean) firm size is larger in the high retail bucket than in the low retail bucket.

In the table footer we formally test for equality between the high retail (Q5) and low retail (Q1) dummy variable coefficients. As the p-values show, all differences except the CAPM beta are statistically significant. We also report the gap in the Q5 and Q1 dummy values, controlling for size by including dummy variables for five size quintiles. The relationships described above continue to hold, with the exception of the difference in B/M ratio that is no longer significant, and past returns which switches sign. Overall, the results in Table 2 establish firm age and nominal price as important determinants of retail trading interest, reflecting the findings in Kumar and Lee (2006) and Balasubramaniam et al. (2023). That said, there are no substantial differences in the standard measures of risk or valuation: CAPM beta and the book-to-market ratio.

In the subsequent Table 3, however, we document substantial differences in valuation and valuation uncertainty metrics across the retail sort, establishing our first main empirical finding: retail investors tend to more heavily trade stocks that are harder to value. For ease of interpretation we winsorize all measures at the 1% level, and then transform into z-scores, meaning we subtract their mean and divide by their standard deviation (see Appendix A.3 for more detail on how these variables are constructed).

The first dimension of difficulty to value is the duration of cash-flows. In the leftmost column of Table 3 we report averages of proxy for cash-flow duration (CF) constructed after Gormsen and Lazarus (2023) and find that high retail stocks tend to have longer duration cash-flows. Also consistent with high retail stocks being harder to value, high retail stocks have a relatively larger share of their value in intangibles, specifically, they have more intangible capital (K_{Int}), as measured in the Peters and Taylor Total Q dataset (Peters and Taylor, 2017). High retail firms also have more valuable patents, relative to their total market value where the value of patents, PAT, is from

(Kogan et al., 2017).[6]

Three existing composite proxies of hard-to-value confirm the relationship between retail trading intensity. PC_{HTV}, constructed in Ben-David et al. (2023), is defined as the first principal component of turnover, analyst forecast dispersion, and idiosyncratic volatility. Consistent with our other results, we find that PC_{HTV} is monotonically increasing from the low to high retail portfolio.[7] High retail stocks have higher valuation uncertainty, denoted VU, constructed in Golubov and Konstantinidi (2021) as the dispersion in expected stock prices given by different accounting-based valuation methods. High retail stocks carry higher mispricing scores constructed from data in Stambaugh and Yuan (2017) who document the average ranking of individual stocks in eleven anomaly portfolios.[8]

Finally, we document the behavior of two return-based proxies. Id. Vol. refers to idiosyncratic volatility as constructed in Kumar (2009): the monthly variance of returns net of the Fama-French 4-factor model. Consistent with this measure increasing in retail trading intensity, we show that high retail stocks have high idiosyncratic volatility in Table A13.[9] Finally, we use the one-month lottery return measure, meaning the average of the five largest daily returns as constructed in Bali et al. (2017) and again find that the proxy for hard-to-value increases in retail trading intensity.

Just like in the prior table we test formally for the equality of coefficient estimates of high and low retail trading intensity quintiles. In all cases the differences across the retail sort are statistically significant. Also mirroring the prior table, we test for the equality of the Q5 and Q1 dummy coefficients controlling for size by including five size dummy variables, as well as a litany of other control variables.[10] In all cases, the differences between Q5 and Q1 are statistically significant and

[6]To compute this metric, we sum the total real dollars of patents over the previous 5 years and divide by the firm's real market capitalization at the end of the focal year.

[7]The relationship between the HTV score of Ben-David et al. (2023) and retail trading intensity is consistent with high retail stocks having higher overall turnover in Table 1, and, looking ahead, with more dispersion in analyst forecasts in Table 4, and higher idiosyncratic volatility in Table A13.

[8]The mispricing score of Stambaugh and Yuan (2017) ranges from 0 to 100, with values near zero or one hundred denoting under- and over-pricing. We construct a mispricing score by calculating the absolute distance of the mispricing measure relative to 50 i.e., identifying stocks that tend to be in extreme (either long or short) anomaly portfolios.

[9]Kumar (2009) finds that overconfidence and the disposition effect are stronger in hard-to-value stocks, defined as those with high idiosyncratic volatility, high turnover and less time since listing. Those results are consistent with our findings that high retail stocks are high turnover (Table 1) and relatively young (Table 2.)

[10]We control for nominal share price, the cumulative return from month $t-12$ to $t-2$ (i.e., the returns used to form momentum portfolios), prior month returns, age, market capitalization, book-to-market, gross profit margin, Fama-French 4-factor betas, total return volatility, and effective spread.

quite similar to the estimates without controls.[11]

Overall, the results in Table 3 establish a new fact: stocks with high shares of retail trading tend to be harder to value.

4.3 Robustness and Validation of the Retail Sort

In this section we offer further validation of our claim that retail investors trade hard-to-value stocks. A potential concern is that the results regarding the spread in retail trading intensity in Table 3 reflect an industry tilt. We calculate RSVOL across the Fama-French 49 industries and find industry-level differences in the share of retail trades. The differences in the metrics included in Table 3, however, are robust to controlling for industry. We find that forming the z-scores within the 49 industries only slightly attenuates the differences between high and low retail stocks, suggesting that the retail tilt toward hard-to-value securities is not driven solely by cross-industry differences.

The Online Appendix contains a number of robustness checks for the relationship between hard-to-value proxies and retail trading intensity. In Table A4 we regress a continuous measure of retail intensity on hard-to-value proxies, controlling additionally for firm size, all the control variables from Table 3 and including month and stock fixed effects. Even in this stringent specification, the hard-to-value proxies retain the positive relationship with retail trading intensity. In Table A5 we explore three alternatives to sorting on retail share of trading volume. We sort on retail-initiated turnover (RTO), 1-institutional ownership, and double sort first on overall monthly turnover (MTO) and then on retail share of trading volume. In each column, the left-hand-side variable is the first principal component of the hard-to-value proxies[12]. Across all alternative sorts, we find that higher retail trading intensity is positively associated with the proxies for hard-to-value.

In Section 5.2, we present alternative evidence that high retail stocks are hard to value by showing that analyst forecast errors tend to be larger among stocks with a high share of retail trades. We motivate this analysis within the model presented in Section 2.1: if a stock is hard-

[11] This is just one set of ways to quantify whether a stock is hard to value. For example, one could imagine that firms with fewer comparable companies, or with more business segments (Cohen and Lou, 2012) would also be harder to value. Also see Décaire et al. (2023) on drivers of analyst disagreement. In this paper, we focus on the "accounting" definitions in Table 3, as they fit more cleanly with our analysis on earnings announcements throughout the rest of the paper.

[12] Valuation uncertainty and absolute mispricing scores are excluded from the principal components calculation due to smaller time-series availability across stocks

to-value then the earnings forecast errors ought to be larger, accounting for the number of stocks a given analyst covers, and accounting for a given analyst's skill. This is precisely what we find. In Table 5 we estimate analyst-announcement level regressions of forecast accuracy—normalized by the number of stocks contemporaneously covered by the analyst—on the quintiles of past retail trading intensity. We find that analyst forecast errors are monotonically increasing from the low to high retail portfolios, both for next quarter earnings forecasts, as well as 12-month price forecasts, and even controlling for firm-level and analyst-level fixed effects. These findings again point to difficulty-to-value as a robust descriptor of firms across the retail sort.

Our evidence focuses on the BJZZ proxy for retail trading intensity. As discussed in Section 3.2, we address in multiple ways the various potential biases in that measure. One way we seek to rule out that measurement error is driving the spread in retail trading intensity is demonstrating that other, slower-moving proxies of retail interest are correlated with the BJZZ measure. To this end, the last column of Table 2 shows that high retail trading intensity stocks tend to have lower institutional ownership.

With detailed Form 13F data we can refine this result further: we find that the holdings of small institutional investors resemble more the retail trading tilt, whereas the holdings of large institutional investors are negatively related to retail trading intensity. To document this difference, we sort 13F-filing institutions into quintiles each quarter based on the total value of their equity holdings. Then, treating each group as one large fund, we compute aggregate portfolio tilts, defined as deviations of each group's portfolio weights from market weights. In Appendix A.7, Panel A of Table A6 shows that in the top quintile of institution size, there is a monotonic decreasing relationship between large institutions aggregate portfolio tilts and retail trading intensity. This monotonic relationship also holds for the next two largest quintiles of institution size. Among the two smallest quintiles of institution size, however, we see a tilt toward high retail stocks. Panel B of Table A6 shows similar using data on the holdings of individual active mutual funds, rather than institutions a as a whole.

We interpret these results as further evidence that the retail sort captures an economically meaningful dimension of the cross-section of stocks. Such difference in holdings across institutional investor size could reflect that large institutions' presumptive informational advantage is mitigated among the high retail stocks, or that large institutions face different investment mandates (Ma et al., 2019; Beber et al., 2021) and constraints on owning a large fraction of small stocks (Edmans et al., 2013), mirroring our main argument regarding the distinction of retail and institutional investors.

5 Testing Predictions on Earnings Announcements

Having established retail traders' tendency to favor hard-to-value stocks, we are now in position to test three sets of predictions developed in Section 2.

5.1 Distribution of Standardized Unexpected Earnings

Prediction 1A states that high retail stocks should see more volatile stock returns and fundamental news around earnings announcements.

To quantify the surprise component of earnings news, we use analyst expectations from IBES. Specifically, for our baseline results, we follow DellaVigna and Pollet (2009) and Hartzmark and Shue (2018), defining standardized unexpected earnings (SUE) as:

$$\mathrm{SUE}_{i,t} = \frac{\mathrm{EPS}_{i,t} - \mathrm{E}_{t-1}[\mathrm{EPS}_{i,t}]}{P_{i,t-1}} \tag{3}$$

where $\mathrm{EPS}_{i,t}$ is earnings-per-share from the IBES unadjusted detail file, that is, "street" earnings. This measure is designed to take out the effect of one-time items (similar to EPSFXQ in Compustat, which excludes extraordinary items). The term "unadjusted" means that earnings were not adjusted for stock splits as is done in the standard IBES summary files. We use the unadjusted file because in constructing the adjusted file, IBES rounds estimates and actual earnings to the nearest penny, which can reduce the precision of the earnings surprise measure.[13] $\mathrm{E}_{t-1}[\mathrm{EPS}_{i,t}]$ is the mean estimate of earnings per share across analysts using their respective most recent values before earnings are released (i.e., the last IBES statistical period) and $P_{i,t-1}$ is the last closing price before the earnings announcement.

The results documenting return and earnings volatility are in Table 4. We calculate the standard deviations on the month-retail-quintile level (for our baseline monthly-rebalanced quintiles formed

[13]For more details see `https://wrds-www.wharton.upenn.edu/documents/5/A_Note_on_IBES_Unadjusted_Data_pdf.pdf`

on prior month retail trading intensity). The first three columns of Table 4 show that high retail stocks have systematically higher return volatility associated with earnings announcements, measured over a 1-, 3-, or 5-day window starting with the first day investors could have traded on the earnings information during normal market hours. On the announcement day, high retail stocks see 210bps of extra volatility, relative to stocks in the middle quintile of past retail activity. For reference, the unconditional announcement day volatility is over 8%. We again test formally for the equality of the Q5 and Q1 values in the table footer. We also report the difference in return volatilities when double sorting stocks first on size and then on past retail trading intensity. The gap in announcement return volatility between the top and bottom quintile of past retail trading activity stocks is somewhat reduced when controlling for size, but is still over 3% on the announcement day itself.

Column 4 of Table 4 shows the standard deviation of SUE across the retail sort. Matching our return volatility calculations, we construct month-retail quintile-level measures of SUE volatility. Consistent with Prediction 1A, high retail stocks tend to see more volatile earnings surprises: the gap in SUE volatility between high and low retail stocks is close to one percentage point. In the table footer we again document that the gap remains statistically significant when controlling for firm size.

To establish why the earnings of these stocks are so hard to predict, in columns 5 and 6 of Table 4 we report differences in the systematic and idiosyncratic components of SUE volatility. To decompose earnings news into systematic and idiosyncratic components, we follow the method in Glosten et al. (2021). Namely, we regress firm-level SUE on market-wide value-weighted SUE and SIC-2 industry-wide value-weighted SUE in trailing five year rolling windows. The systematic component of earnings is the predicted value from this regression in the last year of the five year rolling window, while the idiosyncratic component is the residual. Column 5 shows that while high retail stocks have slightly more systematic SUE volatility, the economic magnitude is small. Further, Column 6 shows that the additional SUE volatility among high retail stocks is essentially all driven by the idiosyncratic component of SUE, indicating that the larger SUE volatility is due to information that is specific to these firms, rather than larger exposure to economy-wide news.[14]

[14]Note that the estimates in Columns 5 and 6 do not add up to the estimates in Column 4. This is because the regression is run in 5 year rolling widows, but the fitted values are only calculated using the last year of data in the window. Therefore, the residuals *in this last year* need not be perfectly uncorrelated from the fitted values.

Collectively, the evidence in the first six columns of Table 4 is consistent with prediction 1A: high retail stocks both have more volatile earnings news and more volatile earnings-day stock returns.

A potential alternative explanation for why high retail stocks have larger earnings surprises is that such stocks have lower analyst coverage on average, which leads to less accurate forecasts. However, in the last column we report the number of analysts and, conditional on size, document that analyst coverage is increasing in the retail intensity, consistent with Martineau and Zoican (2019). Finally, column 7 of Table 4 summarizes the dispersion of analyst forecast errors. We calculate the standard deviation of firm-quarter-analyst level forecast errors and normalize them by pre-announcement stock price to account for the generally lower nominal prices of high retail stocks documented in Table 2. We find that among high retail stocks, analyst forecast errors have a higher standard deviation, again in line with Prediction 1A.

Finally, we carry out an abbreviated version of this analysis using alternative metrics to sort stocks on retail intensity: retail-initiated turnover (RTO), 1-Institutional Share, and a double sort on monthly turnover (MTO) and retail share of trading volume. In Online Appendix Table A7 Panel A we confirm that the findings regarding earnings return volatility continue to hold under these alternative sorts.

5.2 Analyst Accuracy

Another alternative explanation for the results in Table 4 is that there is a systematic difference in the skill of analysts who cover high and low retail stocks. If Prediction 1A is correct, however, we should find that among the stocks a given analyst covers, their estimates will be relatively less accurate for those that are hard to value. In this subsection, we directly test this prediction.

Motivated by the model in Appendix A.1, we propose a measure of analyst accuracy that captures the idea that analysts have limited attention, and therefore their signal precision should depend on the number stocks they cover. To quantify this dependence, for each analyst j, we count how many stocks (indexed by i) the analyst is covering at time t: Num. Stocks $\text{Covered}_{i,j,t}$. The simplest way of capturing limited attention is to divide measures of analyst forecast errors by Num. Stocks $\text{Covered}_{i,j,t}$. The intuition is that analysts equally spread their attention over the stocks they cover, and therefore their prediction accuracy is expected to decrease at rate

1/Num. Stocks $\text{Covered}_{i,j,t}$.[15]

We use this logic to construct two measures of scaled analyst inaccuracy. First, we define scaled analyst forecast inaccuracy for earnings per share (EPS) as:

$$\text{Inccuracy}^{\text{earnings}}_{i,j,t} = \frac{|\,\text{E}_{j,t}[\text{EPS}_{i,t}] - \text{EPS}_{i,t}|/\text{Prc}_{i,t}}{\text{Num. Stocks Covered}_{i,j,t}} \tag{4}$$

where $\text{E}_{j,t}[\text{EPS}_{i,t}]$ is analyst j's estimate for stock i's EPS in quarter t, $\text{EPS}_{i,t}$ is the realized EPS in quarter t and $\text{Prc}_{i,t}$ is the last closing price before earnings were released. To avoid using stale forecasts, we only use forecasts from the last IBES statistical period before earnings are actually released. Similarly, we define scaled analyst forecast inaccuracy for prices at $t+12$ as:

$$\text{Inaccuracy}^{\text{price}}_{i,j,t} = \frac{|\,\text{E}_{j,t}[\text{Prc}_{i,t+12}] - \text{Prc}_{i,t+12}|/\text{Prc}_{i,t+12}}{\text{Num. Stocks Covered}_{i,j,t}} \tag{5}$$

In computing $\text{Inaccuracy}^{\text{price}}_{i,j,t}$ we exclusively use 12-month ahead price forecasts, as this horizon is the best-populated in IBES.[16]

We then test whether analysts are relatively less accurate in predicting the earnings of high retail stocks and low retail stocks. Specifically, we estimate the following regression:

$$\begin{aligned}100 \times \text{Inaccuracy}_{i,j,t} = \alpha &+ \beta_1 Q1_{i,t-1} + \beta_2 Q2_{i,t-1} + \beta_4 Q4_{i,t-1} + \beta_5 Q5_{i,t-1} \\ &+ \gamma X_{i,t-1} + \phi_1 \text{Staleness}_{i,j,t} + a_t + b_j + \epsilon_{i,j,t}\end{aligned} \tag{6}$$

where the Qs are the quintiles of retail trading intensity in month $t-1$, the a_t are a set of time fixed effects, and b_j are a set of fixed effects for each analyst. $\text{Staleness}_{i,j,t}$ is the time (in days) between the IBES statistical period and the date the forecast was made – capturing the staleness of a given forecast.

If high retail stocks are harder to value, one might expect analysts' accuracy to be relatively lower for such stocks, conditional on how many stocks they cover, their skill, and the nature of the firms they're covering. Therefore, we expect β_5 to be positive, and β_1 to be negative. One would also expect ϕ_1 to be positive, as more stale forecasts are likely less accurate on average. There are

[15] See Kacperczyk et al. (2016) for a conditions under which this relationship holds exactly.

[16] Note that there are many potentially stale price forecasts (because price forecasts are not associated with a particular event like earnings forecasts), so we discard any forecast made 90 days before the associated IBES statistical period – which correspond to each calendar month.

many other reasons that analysts may produce inaccurate forecasts, including the nature of the firm, which we aim to capture with a set of control variables $X_{i,t-1}$. This includes essentially all the variables in Tables 2, 3, and A13 which we show are correlated with retail trading intensity.[17] Standard errors are triple clustered at the stock, analyst and time level.

Table 5 contains the results. In columns 1-3, the left-hand-side variable is $100 \times \text{Inaccuracy}^{\text{earnings}}_{i,j,t}$. The unit of observation is analyst-stock-quarter to match the frequency of earnings announcements. Column 1 shows that analysts' relative inaccuracy is significantly higher for high retail stocks than low retail stocks. In fact, the relationship between forecast inaccuracy and quintile of retail trading intensity is monotonic. The mean of $100 \times \text{Inaccuracy}^{earnings}_{i,j,t}$ is 0.058, so the difference in accuracy between the top and the bottom retail quintiles is more than two times the mean.

Column 2 adds analyst fixed effects to account for heterogenity in analyst skill. The point estimates are hardly changed by including these fixed effects, suggesting differences in analyst skill are not a key driver of these results. Finally, column 3 adds a large suite of variables correlated with retail trading activity. While including all these firm level controls does shrink the estimated magnitudes – the difference between the 1st and 5th quintile is still statistically significant, and is on par with the magnitude of the unconditional mean.

In columns 4-6, we repeat the analysis for $100 \times \text{Inaccuracy}^{\text{price}}_{i,j,t}$. The unit of observation is analyst-stock-month, to match the frequency of IBES statistical periods associated with price forecasts. Column 4 shows a large difference in price forecast accuracy for high and low retail stocks. The mean of $100 \times \text{Inaccuracy}^{\text{price}}_{i,j,t}$ is 8.06, meaning the difference between the top and bottom quintiles of retail activity is economically large. In column 5, we include analyst fixed effects, which shrink the point estimates by slightly less than a factor of 2, but the difference between the high and low retail quintiles remains statistically significant. Finally in column 6, we also include all the firm-level controls, and again the relationship between relative inaccuracy is monotonically increasing from low to high retail.

Overall, the results in Table 5 bolster Prediction 1A. Specifically, these results allay concerns that selection on analyst quality drives the relationship between the standard deviation of earnings surprises and retail trading intensity. Further, these effects are economically large – on par with

[17] We omit cashflow duration (Gormsen and Lazarus, 2021), valuation uncertainty (Golubov and Konstantinidi, 2021) and mispricing (Stambaugh and Yuan, 2017) because that would dramatically shrink our sample of observations based on data availability.

the magnitude of unconditional average inaccuracy. More importantly, these estimates provide additional evidence that high retail stocks are hard to value. Even when conditioning on quality through the analyst-level fixed effects, analysts tend to be significantly less accurate in forecasting the earnings of high retail stocks – consistent with their fundamentals being hard to forecast as posited in Prediction 1A.

5.3 Return Sensitivity to Earnings Surprises

Prediction 1B states that high retail stocks should respond less to earnings news than low retail stocks. The logic is that, by nature of having longer duration cashflows and more valuation uncertainly, today's fundamental news is likely less important for today's price. To quantify this relationship, we follow Kothari and Sloan (1992) and estimate earnings response regressions of the form

$$r^i_{t,t+n} = \alpha + \beta \text{SUE}_{i,t} + \gamma X_{i,t} + \phi_t + \psi_i + \epsilon_{i,t}, \tag{7}$$

where $r^i_{t,t+n}$ is the cumulative market-adjusted return from the first day investors could trade on earnings information to n days later.[18] We include both firm and year-quarter fixed effects. Controls in $X_{i,t}$ include a variety of factors known to be correlated with retail activity: nominal share price, the cumulative return from month $t-12$ to $t-2$ (i.e., the returns used to form momentum portfolios), prior month returns, age, market capitalization, book-to-market, gross profit margin, Fama-French 4-factor betas, total return volatility, and effective spread.

We are interested in how the earnings response coefficient (β in Equation (7)) varies across the retail sort, so we interact $\text{SUE}_{i,t}$ with dummy variables for each quintile of retail trading intensity in the month before the earnings announcement. The omitted group is the middle bucket of retail activity. Table 6 contains the results. The first row shows that, consistent with prior work in Kothari and Sloan (1992), SUE is positively related to earnings-day returns. The four interaction terms of RSVOL quintiles and SUE show that high retail stocks respond less to earnings innovations, while low retail stocks respond more to earnings innovations than the average stock. The gap in this sensitivity to fundamental news is large: the difference in coefficients on $\text{SUE} \times Q5$ and $\text{SUE} \times Q1$ is over .6, compared to an unconditional effect of just over 1. In the second set of three columns

[18]Following Campbell et al. (2001), market-adjusted returns are defined as the difference between firm i's return and the market factor from Ken French's data library.

we control for a litany of firm characteristics listed in the above paragraph. The weaker sensitivity of high retail stocks to earnings surprises is left virtually unchanged.

A potential concern with the results in Table 6 is that high retail stocks are not less responsive, but just respond to news more slowly. This aspect would be consistent with the results in Luo et al. (2021) that high retail stocks have a stronger post-earnings announcement drift. Columns 2, 3, 5 and 6 show, however, that the differential response of high retail stocks to earnings news is of roughly constant magnitude over horizons of up to 4 days after the announcement. This stability of coefficient across return horizons suggests that our results are not driven by high retail stocks responding more sluggishly to news.

As discussed above, a number of the characteristics that vary across retail-sorted portfolios in part reflect a size effect. This regularity implies another potential concern with the results in Table 6: retail investors select into small stocks, and such stocks, for instance, by nature of being less covered by media outlets (Martineau and Mondria, 2022) respond less to earnings news. We demonstrate, however, that the weaker sensitivity of high retail intensity stocks to earnings news is not subsumed by size. In Appendix Table A8 we re-estimate the regression 7 but include dummy variables for quintiles of firm size, as well as their interaction with SUE. As Appendix Table A8 shows, high retail share stocks are less responsive to earnings news across the size distribution, and this difference is statistically significant at the 5% level for all but the smallest size portfolios. In Online Appendix Table A7 Panel B we confirm that the lower sensitivity to earnings news is not specific to sorting on the retail share of trading volume, as it also holds across all our alternative proxies of retail trading intensity.

5.4 Retail Trading around Earnings Announcements

The results in the prior section establish substantial differences in the news and return dynamics across the retail sort. In this section we provide evidence that the differences in return dynamics are at least partly due to proactive decisions on behalf of retail traders as a group. We organize our analysis around testing Prediction 2: high retail stocks should see particularly high trading intensity around earnings announcements, and retail ought to act as liquidity providers to the rest of the market around these events.

In Figure 2 we plot abnormal net retail trading around earnings announcements. Within each retail-sorted quintile, net retail trading volume is defined as the volume of retail buys minus the volume of retail sells, normalized by total volume. We subtract from this measure the quintile-level average over the entire sample to construct an abnormal quantity. In the top left panel we show the average abnormal net retail trading volume in stocks belonging to the top and bottom retail quintile around earnings announcements. As the red line indicates, high retail stocks see substantial abnormal retail buying volume in the run-up to earnings announcements. In the last trading day before the earnings news is released, retail traders net buy orders add up to about .8% of total trading volume, relative to about 17% of total retail-originated volume. In other words, the buy tilt of retail traders is substantial in the run-up to earnings announcements. The bottom left panel cumulates such net abnormal trading volume starting 10 days prior to the announcement, illustrating that the cumulative buying by retail amounts to over 2% of the trading volume in the high retail quintile.

The two panels on the right repeat the above analysis, but normalize net retail trading by total shares outstanding, rather than by trading volume. We again find a substantial increase in the net buying by retail in anticipation of earnings announcements, indicating that the effect in the first two panels was not driven by changes in total volume. We include a regression version of these results in Online Appendix Table A10 to demonstrate that the gaps between low and high retail stocks are statistically significant.

Retail investors being net buyers is suggestive of the complement set—institutional investors—exiting high retail stocks ahead of earnings announcements. Prior work in Di Maggio et al. (2021) documents an unconditional version of this phenomenon and argues it arises because institutional investors want to avoid exposure to extreme returns around earnings announcements, as they can lead to outflows. Our findings document that there is significant cross-sectional variation in this phenomenon across stocks – and specifically that institutional avoidance is focused in high retail stocks. That this effect is concentrated in the retail habitat offers a more fundamental explanation for why institutions tend to exit high-retail stocks ahead of earnings announcements: they understand that such hard-to-value stocks have volatile and idiosyncratic earnings-day returns.

5.5 Retail Trading Imbalances and Returns

The aggregate trading of retail investors around earnings announcements—a gradual build-up of abnormally large positions and a subsequent unwinding of these positions—looks like liquidity provision on their part, as the retail traders end up holding a relatively large portion of the announcement risk. The average dynamics around the scheduled events, however, are not on their own enough to conclude that retail traders are actually engaging in liquidity provision. For example, it could equally be the case that retail investors are net buyers because they are on average overly enthusiastic about upcoming announcements.

In order to provide evidence that retail traders are providing liquidity, particularly in the high retail share, hard-to-value stocks, and particularly around earnings announcements, we need to establish the relationship between retail order imbalance and returns. We do so by adapting the framework in Kaniel et al. (2008) to our particular setting with a focus on the cross-section of stocks.

Specifically, we construct marketable retail order imbalance, based on volume – $\text{mroibvol}_{i,t}$ – a firm-week level measure of retail order imbalance:

$$\text{mroibvol}_{i,t} = \frac{\text{RBuy}_{i,t} - \text{RSell}_{i,t}}{\text{RBuy}_{i,t} + \text{RSell}_{i,t}}. \tag{8}$$

From this measure we construct weekly mroibvol quintiles and we refer to the week in which the mroibvol quintiles are calculated as the “focal week”. We then regress excess returns in the weeks surrounding the focal week on said dummy variables. Specifically, we estimate:

$$rx_{i,t-\tau} = a + \sum_{k=1}^{5} b_k 1_{i \in Qmroibvol_{k,t}} + \epsilon_{i,t} \tag{9}$$

where t is the focal week and τ is either -1, 0 or 1. In words, the left-hand-side of Equation 9 is a weekly excess return either in the week when the mroibvol imbalance was calculated, or in the preceding or succeeding week. We estimate these regressions separately for all stocks, and for stocks in the low and high retail trading intensity buckets. We calculate excess returns with respect to an equal-weighted return of all stocks in the sample in a given week. Further, we separately estimate these regressions in all weeks, and only in weeks where the given stock has an earnings announcement.

Table 7 contains the results. Panel A uses all weeks for the analysis, while Panel B restricts to focal weeks that contain an earnings announcement. In all columns, *mroibvol* is measured in the focal week, meaning week $\tau = 0$. Returns are measured in the week indicated in the table header.

The first set of three columns focuses on week -1. The constant is .14 and the coefficient on retail sells is .20, indicating that stocks most sold by retail in week 0 saw positive returns relative to the average in the prior week; the coefficient on retail buys is $-.63$ meaning that the stocks that saw the most retail buying in week 0 had large negative returns in the prior week. Therefore the first column indicates that retail behaves in a contrarian manner with respect to prior week returns, consistent with providing liquidity to the market. In the table footer we report a test of the equality of returns between extreme retail buys and sells. We find that the gap is -83 basis points and statistically significant at the 1% level.

The second and third columns repeat the same analysis, but restrict the sample to low and high retail trading intensity stocks. In both subsets the same conclusion holds: retail traders as a group behave in a contrarian manner with respect to prior week returns. Note, though, that the effect size is more than three times as large for stocks with high past retail trading activity, consistent with heightened retail liquidity provision in this subset of stocks.

The second set of three columns documents the contemporaneous relationship between retail order imbalance and returns. Here the results tend to be quite muted, and the return differential between heavily bought and heavily sold stocks is not statistically significant. Only for the high retail stocks do we see a gap in the returns between extreme retail buys and sells, but the coefficients do not show a consistent pattern.

The third set of three columns in Panel A of Table 7 studies the predictive power of retail imbalances for future returns. Looking across all stocks (Column 7) we see that intense retail buying has predictive power for returns in the same direction in the subsequent week, consistent with the results in Kaniel et al. (2008). Stocks heavily sold by retail see a negative excess return of 8 basis points in the subsequent week while heavily bought stocks see a positive excess return of 10 basis points, relative to the baseline of 1 basis point return. Such a gap in next week returns is particularly pronounced for stocks with a high retail presence: as the final column shows the gap between heavily bought and heavily sold stocks is on average 33 basis points.

Overall, these results emphasize the importance of recognizing cross-sectional heterogeneity in retail trading intensity. The stocks with high retail presence see more contrarian buying after poor returns and stronger return predictability from retail flows.

Panel B of Table 7 repeats the above analysis but restricts to focal weeks when an earnings announcement took place. The overall results are strikingly similar to the ones reported in Panel A: retail buying tends to be contrarian with respect to prior week returns, and retail buying has predictive power over next week returns. The magnitudes of the sensitivities to retail imbalance tend to be marginally larger when focusing on earnings weeks.

As discussed at length in Section 3.2, recent work has documented both noise and bias in the BJZZ algorithm that we use to identify retail trades and assign retail trade direction. Barber et al. (2023), in particular, find that an improved algorithm employing the Lee and Ready (1991) method of assigning trade direction improves accuracy. We implement their algorithm and re-estimate Table 7. The alternative estimates reported in Online Appendix Table A11 are very close to the original values. If anything, the gap in returns we document is slightly stronger. Our interpretation of the findings is that despite the substantial noise in the BJZZ algorithm, there is not much in the way of directional bias with respect to the relationship between retail imbalance and returns.

5.6 Decomposing Return Predictability

The positive predictability of retail order imbalance for future returns documented in Table 7 leaves open at least two possibilities: either retail trades as a group are providing liquidity to the market, or they possess private information or skill. We report the results from one approach that seeks to attribute relative weights to these two possibilities. In Table 8 we follow the methodology of Kaniel et al. (2012) in order to decompose these average returns into liquidity provision, and into private information.

In each day of the sample we estimate a cross-sectional regression of long horizon (60 trading day) cumulative abnormal returns on indicator variables for quintiles of retail order imbalance (mroibvol), formed over the the past ten days, controlling for past cumulative abnormal returns over the past ten days and limiting the sample to stocks not within 20 trading days of an earnings announcement. We then use these estimated coefficients to predict the expected cumulative abnormal return (ECAR) for firms making an earnings announcement on that day.

In the first three columns of Table 8 we regress firm-announcement level cumulative abnormal returns on indicator variables for pre-announcement retail order imbalance. The first column shows that stocks which were heavily bought by retail traders in the run-up to an earnings announcement see returns of 1.15% over the next 61 trading days, starting with the day the announcement content was first tradeable. This effect is stronger for stocks that are in the high retail trading quintile, as shown in the third column: the gap between returns of heavily bought and heavily sold stocks is 2.3%.

The second set of three columns reports the same analysis with ECAR—expected cumulative abnormal return—-as the left hand side variable. ECAR is estimated contemporaneously from the relationship between retail imbalance and future CAR of non-announcing firms. The ECAR returns, therefore, can be interpreted as the current expected level of returns from liquidity provision because they reflect the current relationship between retail imbalances and returns in non-announcing firms. As the fourth column shows, across all stocks the predicted return differential from liquidity provision is about .92%. The estimated return from liquidity provision among the highest retail stocks is 1.37% leaving over .9% of the return differential to private information. In Online Appendix Table A12 we again confirm that the decomposition is quantitatively unchanged using the improved trade signing algorithm advocated by Barber et al. (2023).

In all, we find a stronger return predictability from retail imbalances for stocks that see a higher level of retail trading. Table 8 suggests that about half of this return predictability can be attributed to liquidity provision, which is the economic role we have suggested retail traders take in the stocks they trade most ahead of earnings announcements.

About half of this return predictability, though, can be attributed to private information. A natural question is then: what is the source of this apparent retail informational advantage? For one, it might be that the signals of individual retail investors aggregate up to a precise signal, even if any particular retail investor is not well informed. Among high retail stocks this could be true as institutional investors know these are the stocks where their presumptive informational advantage is weakest and shy away from information production.

Another possibility is that abnormal retail order flow predicts future abnormal flows in the same direction, as shown in Boehmer et al. (2021). This persistence makes betting against retail orders risky, as more orders in the same direction may arrive and force early liquidation at a loss. In that sense, a large directional move on part of retail trades may end up looking like a response to "fundamental" news, as it will not be corrected quickly. This mechanism is illustrated by the model in De Long et al. (1990), where the risk of persistence in noise trader behavior deters rational arbitrageurs from trading against them – essentially allowing noise traders to "create their own space," making it seem as though they have fundamental information.

5.7 Trading Costs around Earnings Announcements

The substantial shifts in trading around earnings announcements as well as the substantial differences in earnings content and volatility across the retail sort suggest that trading costs are also likely related to past retail trading intensity. The direction of the effect presents a microcosm of our overall study, with plausible economic forces in both directions.

On one hand, if retail investors act as pure noise traders, and/or if institutional investors avoid learning about high retail share stocks, one might expect relatively lower transaction costs. Within the classic Kyle (1985) logic discussed in Appendix A.1 larger amounts of noise trading, or a lower signal precision for insiders decrease the market maker's risk of adverse selection and therefore price impact (Kyle's λ).

On the other hand, retail investors are not pure noise traders and as suggested by the return decomposition in Table 8, potentially have information about future fundamentals. Even if retail investors themselves have no information about fundamentals, retail order flow is persistent. This persistence makes providing liquidity to retail trades risky, as it's possible that subsequent retail trades in the same direction will further push prices against the market maker's position. These forces can make providing liquidity to retail investors relatively riskier, and therefore lead to increased trading costs.

We provide estimates of trading costs around earnings announcements in Table 9. Given that there is an unconditional difference in trading costs across the retail sort – with high retail stocks generally having larger effective percentage bid-ask spreads – we demean trading costs within retail quintiles. Given that there are substantial differences in the earnings news volatility across the retail sort, we also control for the SUE surprise quintile within the quarter.

To account for both of these features, we estimate the following regression:

$$\begin{aligned}\text{DM Effective Spread}_{i,t} = \alpha + \beta_1 1_{i\in Q1_{\tau-1}} + \beta_2 1_{i\in Q2_{\tau-1}} + \beta_4 1_{i\in Q4_{\tau-1}} + \beta_5 1_{i\in Q5_{\tau-1}} + \\ \theta_1 1_{i\in \text{Q1SUE}_t} + \theta_2 1_{i\in \text{Q2SUE}_t} + \theta_4 1_{i\in \text{Q4SUE}_t} + \theta_5 1_{i\in \text{Q5SUE}_t} + \gamma X_{i,t} + \phi_t + \psi_i + \epsilon_{i,t}\end{aligned} \tag{10}$$

where DM Effective Spread$_{i,t}$ is the demeaned effective spread, defined as the effective bid-ask spread from the WRDS intraday indicators suite minus the average effective spread for that stock in the month before the earnings announcement. $1_{i\in QkSUE_t}$ is an indicator for whether firm i's SUE is in the kth quintile of SUE among all firms that released earnings that quarter.

The results in Table 9 show that, even conditional on the nature of the earnings news and differences in average trading costs, high retail stocks are especially expensive to trade before, on, and after earnings announcements. In terms of magnitudes, in the pre- and post- earnings period, high retail stocks are about 2 basis points more expensive to trade, while on the earnings day itself they are about 4 basis points more expensive to trade. This 2-4 basis point increase is large relative to the unconditional value-weighted bid-ask spread in 2021, which was 6 basis points.[19]

Overall, the results in Table 9 suggest that market makers are especially concerned about adverse selection risk in high retail stocks around earnings announcements.

5.8 The Earnings Announcer Premium across Retail Portfolios

The results in the prior subsections establish that high retail stocks are less sensitive to earnings news. In this final subsection we show that this gap in terms of sensitivity translates into a return differential in portfolios that take exposure to announcing stocks as a function of their past retail trading intensity. Our analysis is motivated by the finding in Savor and Wilson (2016) that announcing firms outperform those with no scheduled announcements, and that the aggregate announcer portfolio has alpha with respect to the buy-and-hold stock market portfolio. We aim to refine this result and test Prediction 3, which argues that the earnings announcer premium should be lower, or non-existent, among high retail stocks.

To test this hypothesis, in Table 10, we decompose average returns around earnings announce-

[19] While these magnitudes may appear small, these regressions control for a host of firm-level characteristics and fixed effects. Further, because the results are in terms of *demeaned* effective spreads, they also account for the larger average trading costs for high retail firms.

ments into pre- and post- announcement components as a function of size and retail trading intensity.

The first three columns focus on a narrow window: the last trading day before the earnings announcement, and the first trading day on which the announcement could have been traded. The second set of three columns focuses on a 6-day announcement window, containing three trading days prior to the announcement, the day the earnings news could have been first traded on, as well as the next two trading days. In both sets of announcement windows, "Pre" refers to the portion prior to the announcement, and "Post" refers to the portion after the announcement. Each panel restricts the sample to the indicated size quintile, and Q5 is the dummy variable for the 20% of stocks with the highest share of retail trading within that size bucket. All regressions contain month dummies and standard errors are clustered by day and firm.

The first takeaway from Table 10 is the presence of the earnings announcer premium. Specifically, in the first column and the fourth column, the coefficient on the size dummies is always positive, suggesting that, on average, announcing firms have positive returns.

Secondly, Table 10 shows that high retail stocks see lower announcement time returns, in line with Prediction 3. Focusing first on the third column, representing the first trading day on which earnings announcement is tradeable, the bottom panel shows that among stocks in the top quintile of market capitalization i.e., the largest stocks, the average announcement time return is 14 bps. Similarly, the average announcement time returns are 30, 41, 30, and 7 basis points for the remaining size quintiles. In all cases, the announcer premium is considerably smaller for the high retail share stocks—compared to stocks in Q3—and in all cases this difference is statistically significant. In fact, the coefficient on Q5 is in all cases larger than the unconditional return.

The second set of three columns repeats the same analysis over a six-day event window straddling the earnings announcement. Again the same pattern emerges: high retail stocks underperform others in the earnings announcement window, and this gap is present across the size quintiles, representing mostly lower post-announcement returns. Overall, the findings replicate the known result that average stock returns are high around earnings announcements but find a substantial amount of heterogeneity across the retail sort: the announcement risk premium is negligible among high retail stocks.

In Online Appendix Table A14 we confirm that the findings on earnings announcement returns are qualitatively unchanged using the three alternative metrics of retail trading intensity.

Another notable feature of Table 10 is the difference in pre-earnings and post-earnings returns for high and low retail stocks. Specifically, within every size quintile, high retail stocks have higher pre-announcement returns than low retail stocks. High retail stocks, however, have consistently lower post-announcement returns than low retail stocks. We believe these high pre-announcement returns may be compensation for liquidity provision by retail investors. As shown in Figure 2, institutional investors exit ahead of earnings announcements, and thus retail must be compensated for bearing this extra risk. Now, a natural question is why institutional investors seem to be continuously exiting before – and thus why this effect is larger when examining a larger pre-announcement window. One could imagine instead that institutional investors exit all at once on the day before the announcement itself. We believe the strategy of waiting until the last minute is risky, as (1) firms may not always release earnings when expected, (2) earnings information may leak into the market before the earnings announcement itself, and (3) there may be very little liquidity in the market at that time. This slow exit by institutional investors may also reflect inventory management by market makers. As discussed in Johnson and So (2018), market makers, by virtue of holding inventory, are naturally long – and may want to reduce this risk ahead of earnings announcements. This may be especially salient for high-retail stocks, where the announcements themselves tend to be the most volatile. As a result, market makers may reduce their inventory in high retail stocks by slowly selling out of their positions ahead of the announcements themselves.

In sum, the results in Table 10 are consistent with Prediction 3: high retail stocks do not earn the earnings-announcer premium. Savor and Wilson (2016) argue the earnings announcer premium is compensation for exposure to systematic news. And, as shown in Table 4, high retail stocks' SUE is mostly composed of idiosyncratic news. Therefore, one explanation for the findings in Table 10 is that high retail stocks do not earn the announcer premium because their SUE is mostly composed of idiosyncratic information.

6 Conclusion

In this paper, we establish a new fact: retail investors tend to favor trading stocks which are hard to value. Consistent with this cross-sectional regularity, such stocks have more volatile realizations of both fundamental news and earnings-day returns. Further, these stocks tend to respond less to earnings news of a given size, and are relatively more expensive to trade around earnings announcements.

We additionally document how retail investors trade around earnings announcements. Retail are abnormally active in the pre-earnings announcement period, acting as net buyers from institutional investors, particularly in the stocks they favor generally. Intense buying by retail in the run-up to announcements predicts positive excess post-announcement returns, particularly for stocks with a large retail presence.

Finally, we link the fact that retail investors favor hard to value stocks to the earnings announcer premium. Past literature has argued that this premium is earned as compensation for exposure to the systematic risk contained in earnings news. We find that high retail stocks have a small systematic component in their earnings news and that any news about these firms is hard to interpret. So, consistent with the systematic risk-based explanation of the earnings announcer premium, it is not earned in high retail stocks.

Overall, our findings document a new dimension of investor heterogeneity. Retail investors have a comparative advantage relative to institutional investors in trading hard-to-value stocks. Further, we find that stocks with significant retail trading activity have low institutional ownership, suggesting retail investors also have a comparative advantage in holding such stocks. This pattern is especially stark when comparing retail to large investment managers, suggesting institutional constraints are an important determinant of aggregate risk sharing. Therefore, our results speak to an novel dimension of cross-sectional heterogeneity in which groups of investors bear different types of risk.

7 Figures

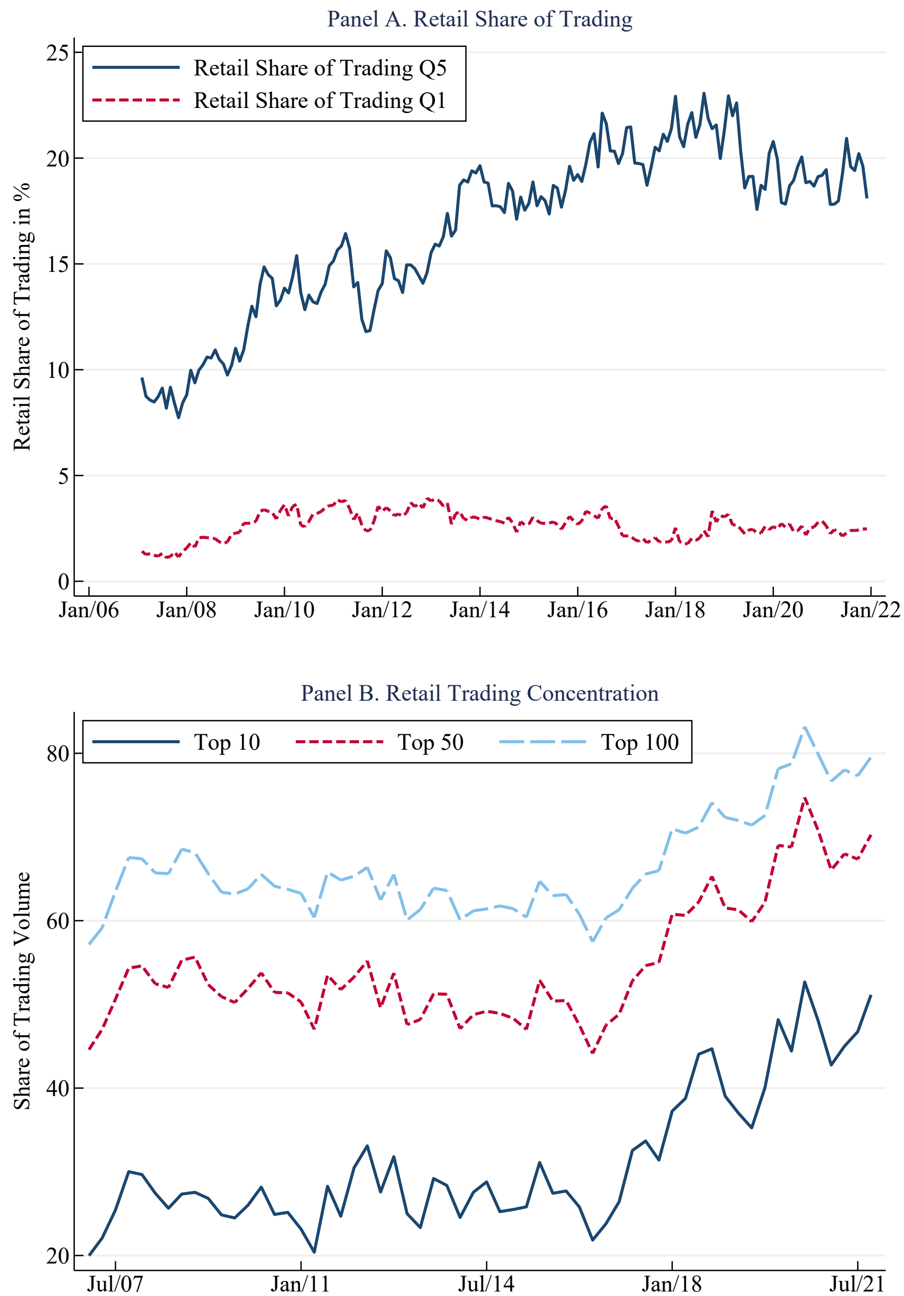

Figure 1: Retail share of trading volume and retail trading concentration. Panel A shows the average retail share of trading volume in the top and bottom quintiles sorted on previous month's retail trading intensity. Panel B shows the cumulative share of total retail dollar volume stemming from the top 10, 50, and 100 stocks sorted by retail dollar volume.

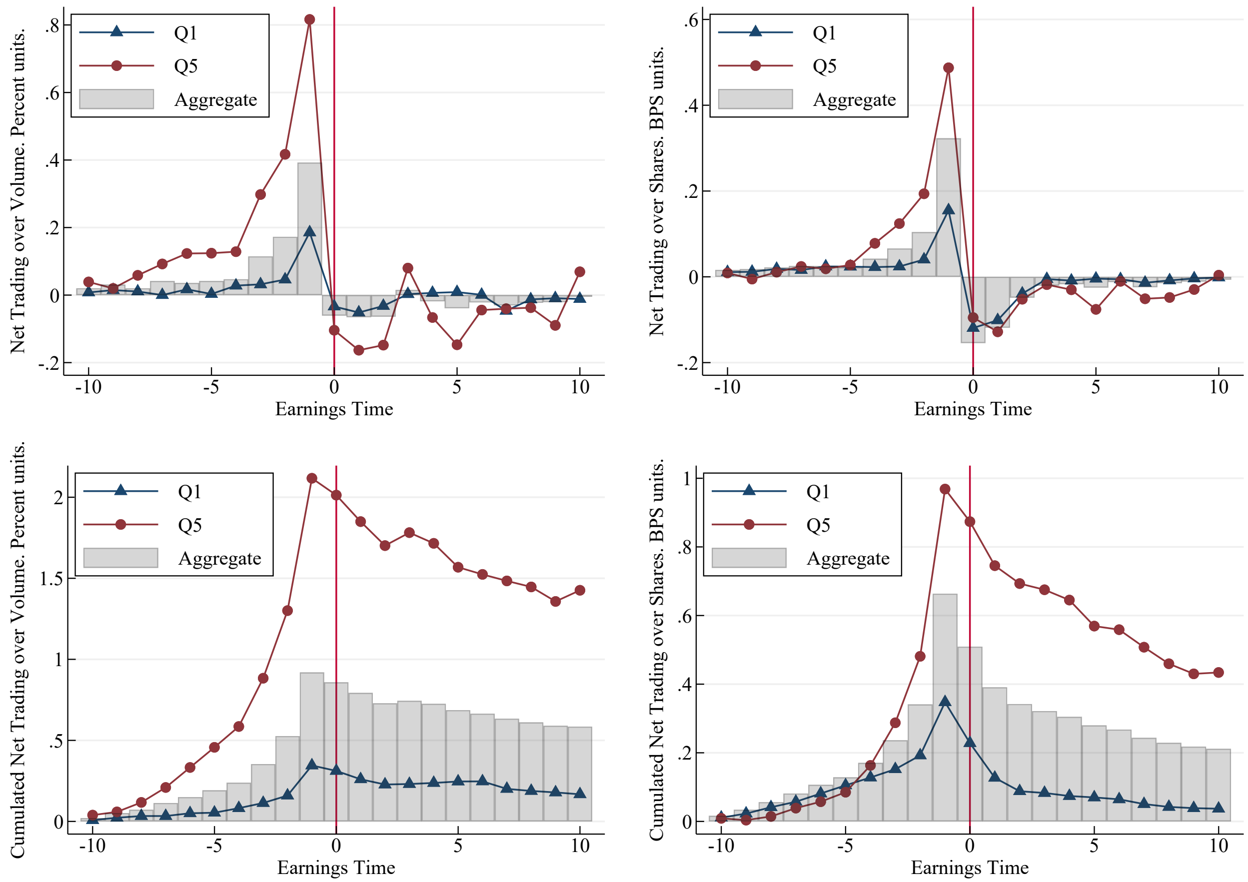

Figure 2: Abnormal net trading around earnings announcements. Daily net trading (retail-initiated buys minus retail-initiated sells, measured in shares), normalized by aggregate trading volume or shares outstanding. We subtract out the unconditional means in respective series to construct abnormal measure and take an equal-weighted average within each quintile. Q1 represents the bottom quintile of retail intensity, while Q5 represents the top quintile. Bottom panels cumulate the values in top panels staring at time -10 relative to earnings announcement day at time 0.

8 Tables

	Retail Trading			Turnover			Retail-initiated TO		
	All	Small	Large	All	Small	Large	All	Small	Large
Low	-2.42***	-2.88***	-2.31***	-0.82***	-1.45***	0.33	-0.42***	-0.20***	-0.40***
	(-66.58)	(-64.07)	(-69.32)	(-3.31)	(-5.12)	(0.67)	(-37.26)	(-5.48)	(-21.73)
2	-1.29***	-1.36***	-1.30***	-0.52***	-0.77***	-0.65**	-0.24***	-0.12***	-0.24***
	(-65.83)	(-52.59)	(-57.40)	(-3.12)	(-3.51)	(-2.02)	(-30.15)	(-14.37)	(-18.59)
4	2.51***	2.74***	2.50***	1.80***	1.58***	2.97***	0.57***	0.33***	0.64***
	(52.95)	(48.60)	(48.90)	(6.45)	(7.43)	(4.42)	(23.32)	(17.98)	(13.62)
High	10.84***	12.46***	8.21***	7.10***	9.46***	26.50***	2.91***	2.52***	4.98***
	(49.44)	(50.35)	(25.51)	(7.84)	(12.00)	(8.96)	(20.03)	(17.84)	(10.52)
Average	6.37	12.38	4.51	20.68	14.30	22.06	1.32	1.90	1.02
Q5-Q1	13.26	15.34	10.52	7.93	10.91	26.17	3.34	2.72	5.37
p(Q1=Q5)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N	462,558	92,581	92,439	462,558	92,581	92,439	462,558	92,581	92,439
R^2	0.72	0.54	0.77	0.02	0.04	0.12	0.21	0.10	0.38

Table 1: Trading in five retail share of trading sorted portfolios. Firm-month level regressions of retail share of trading, turnover, and retail-initiated turnover on prior month retail trading intensity. Low refers to the quintile with least retail-initiated trading in the prior month; high refers to the quintile with most retail-initiated trading in the prior month. Small and Large refer to the first and fifth quintile in terms of firm size, respectively. The bottom part of the table reports the average value of the left-hand-side variable, the gap between Q5 and Q1, and the associated p value. Standard errors clustered on the firm and month level.

	Cap	Age	Prc	Past R	B/M	E/P	β_{CAPM}	100-Inst.
Low	0.53***	-0.69	1.73	-1.89***	0.01	0.01***	-0.05***	-5.74***
	(38.24)	(-1.31)	(1.55)	(-4.27)	(1.14)	(9.28)	(-6.48)	(-13.41)
2	0.64***	0.24	3.42***	-0.59*	-0.03***	0.01***	-0.02***	-4.77***
	(39.53)	(0.71)	(4.86)	(-1.88)	(-4.53)	(11.81)	(-4.93)	(-19.51)
4	-0.73***	-1.71***	-9.06***	-1.01**	0.09***	-0.03***	0.02***	8.49***
	(-65.26)	(-4.18)	(-9.74)	(-2.11)	(8.39)	(-17.65)	(3.22)	(26.32)
High	-1.16***	-7.87***	-27.62***	-7.58***	0.23***	-0.15***	-0.05***	31.49***
	(-111.65)	(-16.15)	(-19.49)	(-8.76)	(9.19)	(-23.90)	(-3.68)	(49.48)
Average	7.26	20.67	34.78	9.47	0.65	0.00	1.11	31.04
Q5-Q1	-1.69	-7.18	-29.35	-5.68	0.22	-0.16	0.00	37.24
p(Q1=Q5)	0.00	0.00	0.00	0.00	0.00	0.00	0.92	0.00
Q5-Q1, size	-0.03	-1.39	-6.61	10.07	-0.03	-0.10	0.24	19.86
p(Q1=Q5), size	0.00	0.02	0.00	0.00	0.29	0.00	0.00	0.00
N	416,432	416,432	416,432	416,432	416,432	416,432	416,432	416,432
R^2		0.03	0.08	0.00	0.02	0.11	0.00	0.26

Table 2: Fundamentals in five retail share of trading sorted portfolios. Firm-month level regressions on dummy variables representing retail trading intensity quintiles formed in the prior month. Cap is market cap; Age is time since listing; Prc is nominal price; Past R is the returns from month $t = -12$ to $t = -2$ i.e., the returns used to form momentum portfolios (Jegadeesh and Titman (1993)); B/M is book-to-market; E/P is the earnings-to-price ratio; β_{CAPM} is the market beta computed over the previous 252 trading days; Inst. is institutional ownership from 13F data and 100-Inst. is therefore an alternative proxy for retail ownership. The first column estimates a median regression. The bottom part of the table reports the average value of the left-hand-side variable, the gap between Q5 and Q1, and the associated p value, as well as the Q5-Q1 gap controlling for firm size. Standard errors clustered on the firm and month level.

	CF	K_{Int}	PAT	PC_{HTV}	VU	Mispric.	Id. Vol.	Lottery
Low	-0.10***	-0.05***	-0.17***	-0.22***	-0.11***	-0.07***	-0.23***	-0.17***
	(-3.87)	(-4.02)	(-7.69)	(-12.40)	(-4.42)	(-4.21)	(-20.60)	(-18.00)
2	-0.08***	-0.05***	-0.07***	-0.13***	-0.09***	-0.03***	-0.16***	-0.12***
	(-4.95)	(-6.98)	(-4.78)	(-11.00)	(-5.80)	(-2.65)	(-21.42)	(-18.14)
4	0.12***	0.18***	0.08***	0.25***	0.25***	0.04***	0.30***	0.22***
	(5.98)	(12.25)	(4.00)	(13.61)	(12.97)	(2.72)	(24.23)	(20.72)
High	0.26***	0.56***	-0.01	0.76***	0.77***	0.20***	1.00***	0.74***
	(8.62)	(18.46)	(-0.23)	(22.25)	(24.50)	(8.59)	(40.41)	(33.34)
Q5-Q1	0.36	0.61	0.16	0.98	0.87	0.27	1.23	0.91
p(Q1=Q5)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q5-Q1, controls	0.23	0.19	0.29	0.75	0.18	0.26	0.77	0.62
p(Q1=Q5), controls	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N	348,122	460,598	462,558	324,543	216,981	256,831	462,558	462,558
R^2	0.02	0.05	0.01	0.12	0.10	0.01	0.20	0.11

Table 3: Valuation across five retail share of trading sorted portfolios. Firm-month level regressions on dummy variables representing retail trading intensity quintiles formed in the prior month. All valuation metrics transformed into z scores for ease of interpretation. CF is cashflow duration, computed after Gormsen and Lazarus (2023); K_{Int} is a measure of intangible capital, normalized by market capitalization (data from Peters and Taylor (2017)); PAT is the real market value of patents over the past five years (data from Kogan et al. (2017)) divided by market capitalization; PC_{HTV} is a composite hard-to-value measure from Ben-David et al. (2023); VU is valuation uncertainty from Golubov and Konstantinidi (2021); Mispric. is a composite mispricing score from Stambaugh and Yuan (2017); Id. Vol. is idiosyncratic volatility as constructed in Kumar (2009); Lottery is the 1-month lottery return as constructed in Bali et al. (2017). The bottom part of the table reports the gap between Q5 and Q1, and the associated p value, as well as the Q5-Q1 gap controlling for characteristics described in Section 4.2. Standard errors clustered on the firm and month level.

	SD(Returns)			SD(SUE)			Analysts	
	(0, 0)	(0, 2)	(0, 4)	Full	System.	Idiosyn.	SD	Number
Low	-1.17***	-1.44***	-1.66***	-0.37***	0.02	-0.47***	-0.10***	-0.65***
	(-8.58)	(-10.34)	(-10.91)	(-9.31)	(1.11)	(-4.72)	(-10.67)	(-3.15)
2	-0.66***	-0.88***	-1.04***	-0.32***	0.00	-0.38***	-0.08***	0.20
	(-5.11)	(-6.47)	(-6.58)	(-7.26)	(0.29)	(-4.54)	(-11.02)	(1.58)
4	1.12***	1.75***	2.01***	0.61***	0.01	1.23***	0.21***	-0.86***
	(8.54)	(10.38)	(9.98)	(12.71)	(0.52)	(8.11)	(13.43)	(-5.04)
High	2.10***	3.36***	4.26***	1.95***	0.02	3.21***	1.01***	-3.71***
	(13.28)	(16.43)	(16.59)	(31.74)	(1.08)	(10.30)	(16.22)	(-15.98)
Average	8.42	9.99	10.86	1.60	0.29	2.24	0.42	9.15
Q5-Q1	3.28	4.81	5.91	2.32	0.00	3.67	1.11	-3.06
p(Q1=Q5)	0.00	0.00	0.00	0.00	0.87	0.00	0.00	0.00
Q5-Q1, size	3.24	4.23	4.82	1.14	0.01	1.99	0.76	1.53
p(Q1=Q5), size	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N	895	895	895	895	895	895	135,026	135,026
R^2	0.23	0.29	0.30	0.53	0.00	0.16	0.05	0.04

Table 4: Earnings Announcement Summary Statistics. Quarterly earnings announcements. In the first six columns we report retail quintile-month level regressions. SD(Returns) refers to the standard deviation of announcement returns, measured starting on day 0, the first day on which the announcement is tradeable. SD(SUE) refers to the standard deviation of SUE, measured separately for the full SUE, as well as for the systematic and idiosyncratic components of SUE. Standard deviations constructed on the retail quintile-month level. In the last two columns, SD and Number refer to the standard deviation of analyst forecasts and the number of analysts, respectively. These variables are measured on the stock-announcement level. The bottom part of the table reports the average value of the left-hand-side variable, the gap between Q5 and Q1, and the associated p value, as well as the Q5-Q1 gap controlling for firm size. Standard errors clustered on the quintile/firm and quarter level.

	Forecasts of Next Quarter Earnings			Forecasts of Price in 12 Months		
	(1)	(2)	(3)	(4)	(5)	(6)
Low Retail	-0.0138***	-0.0166***	-0.00947***	-1.438***	-0.974***	-0.824***
	(0.003)	(0.004)	(0.002)	(0.218)	(0.175)	(0.193)
2	-0.0127***	-0.0120***	-0.00557***	-0.982***	-0.598***	-0.423***
	(0.003)	(0.003)	(0.002)	(0.145)	(0.122)	(0.137)
4	0.0334***	0.0315***	0.0126***	2.825***	1.965***	1.308***
	(0.007)	(0.007)	(0.004)	(0.348)	(0.264)	(0.205)
High Retail	0.132***	0.118***	0.0421***	11.43***	6.356***	3.526***
	(0.020)	(0.021)	(0.008)	(0.914)	(0.529)	(0.450)
Observations	1,024,786	1,024,339	1,024,339	1,939,687	1,939,631	1,939,631
R-squared	0.059	0.128	0.179	0.097	0.28	0.291
Fixed Effects	YQ	YQ/Analyst	YQ/Analyst	YM	YM/Analyst	YM/Analyst
Firm-Level Controls	NO	NO	YES	NO	NO	YES
Clustering		Stk/YQ/Analyst			Stk/YM/Analyst	

Table 5: Individual analyst inaccuracy and retail trading intensity. Quarter/month-firm level regressions. Left-hand-side variables are measures of analyst inaccuracy, scaled by the number of stocks the analyst covers. The firm-level controls are market capitalization, firm age, returns from month $t-12$ to month $t-2$, the MAX factor of Bali et al. (2017) (a measure of lottery demand), book-to-market, earnings-to-price, market beta, standard deviation of daily returns, intraday idiosyncratic volatility computed from trades, Kyle's lambda, the effective bid-ask spread, intangible capital divided by market capitalization and the dollar value of patents divided by market capitalization. The unit of observation for columns 1-3 is stock-quarter-analyst, while the unit of observation for columns 4-6 is stock-month-analyst. All left-hand-side variables have been multiplied by 100 to ease the interpretation of the associated coefficients. Standard errors are triple clustered at the stock, analyst and time levels.

	Standardized Unexpected Earnings					
	(0, 0)	(0, 2)	(0, 4)	(0, 0)	(0, 2)	(0, 4)
SUE	1.06***	1.16***	1.17***	1.09***	1.19***	1.23***
	(14.79)	(11.25)	(11.66)	(14.17)	(10.99)	(11.48)
SUE x Q1	0.16	0.18	0.20	0.21*	0.24	0.26*
	(1.93)	(1.47)	(1.72)	(2.35)	(1.88)	(2.03)
SUE x Q2	0.34***	0.35**	0.39**	0.39***	0.43***	0.45***
	(4.07)	(3.00)	(3.02)	(4.19)	(3.51)	(3.46)
SUE x Q4	-0.18*	-0.14	-0.10	-0.19*	-0.13	-0.09
	(-2.57)	(-1.40)	(-0.96)	(-2.60)	(-1.21)	(-0.80)
SUE x Q5	-0.47***	-0.47***	-0.45***	-0.43***	-0.40***	-0.40***
	(-6.85)	(-4.78)	(-4.69)	(-5.85)	(-3.98)	(-3.98)
Controls				Yes	Yes	Yes
N	148,932	148,932	148,932	138,623	138,623	138,623
R^2	0.09	0.09	0.09	0.10	0.10	0.10

Table 6: Post-announcement return sensitivity to realized standardized earnings surprise. Firm-announcement level regressions of post-announcement returns on standardized unexpected earnings. Post-announcement return period indicated in column header. 0 refers to the first day announcement information is tradeable during normal market hours. SUE is standardized unexpected earnings, defined in Equation 3 and Qk is an indicator variable for whether stock i was in retail intensity quntile k at the end of the month before the earnings announcement. SUE and returns winsorized at the 1st and 99th percentile. All regressions control for stock and month dummies. Additional controls include nominal price, returns from month $t-12$ to $t-2$, time since listing, market capitalization, book-to-market, gross profit margin, book long-term leverage, MAX (lottery demand) and month $t-1$ returns and betas on the three Fama-French factors as well as the momentum factor, idiosyncratic volatility and total volatility, all computed over the past 12 months. Standard errors clustered on the stock and month level.

.

Panel A.

	Week -1			Week 0			Week 1		
	All	Low	High	All	Low	High	All	Low	High
Retail Sells	0.20***	0.20***	-0.07	-0.28***	-0.06	-0.93***	-0.08***	-0.06*	-0.09
	(6.25)	(5.60)	(-0.85)	(-7.88)	(-1.58)	(-10.93)	(-2.80)	(-1.65)	(-1.19)
2	0.05***	0.11***	-0.26***	-0.24***	-0.02	-0.93***	-0.00	0.00	-0.02
	(2.79)	(4.53)	(-4.20)	(-10.88)	(-0.67)	(-15.26)	(-0.26)	(0.16)	(-0.36)
4	-0.24***	-0.11***	-0.56***	-0.11***	-0.09***	0.10	0.01	0.07***	-0.03
	(-12.60)	(-4.49)	(-9.00)	(-5.64)	(-3.47)	(1.55)	(0.92)	(2.79)	(-0.58)
Retail Buys	-0.63***	-0.26***	-1.38***	-0.38***	-0.16***	-0.50***	0.10***	0.05	0.24***
	(-21.42)	(-8.08)	(-17.16)	(-12.61)	(-4.94)	(-6.19)	(4.13)	(1.57)	(3.19)
Constant	0.14***	0.08**	0.36***	0.22***	0.13***	0.33***	0.01	0.05	-0.16**
	(9.31)	(2.38)	(4.54)	(14.00)	(4.01)	(4.31)	(0.44)	(1.62)	(-2.17)
Sells-Buys	-0.83	-0.45	-1.31	-0.11	-0.10	0.44	0.18	0.10	0.33
Sells=Buys p	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N	1,907,139	382,104	361,259	1,907,139	382,104	361,259	1,907,139	382,104	361,259
R^2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Panel B.

	Week -1			Week 0			Week 1		
	All	Low	High	All	Low	High	All	Low	High
Retail Sells	0.21***	0.12	0.17	-0.44***	-0.21	-1.04***	-0.17**	-0.04	-0.25
	(3.15)	(1.40)	(0.95)	(-3.40)	(-1.17)	(-3.70)	(-2.19)	(-0.44)	(-1.28)
2	0.09*	0.01	-0.03	-0.46***	-0.38**	-1.04***	-0.07	0.04	-0.03
	(1.94)	(0.08)	(-0.23)	(-4.46)	(-2.37)	(-4.08)	(-1.45)	(0.59)	(-0.19)
4	-0.18***	-0.10	-0.27	-0.75***	-0.29**	-1.18***	-0.05	-0.03	-0.04
	(-3.55)	(-1.46)	(-1.57)	(-8.66)	(-2.00)	(-4.17)	(-1.12)	(-0.44)	(-0.26)
Retail Buys	-0.74***	-0.36***	-1.28***	-1.73***	-0.73***	-1.84***	-0.00	0.01	0.04
	(-10.58)	(-4.48)	(-6.69)	(-14.84)	(-4.86)	(-6.26)	(-0.05)	(0.12)	(0.22)
Constant	0.18***	0.21***	0.18	0.55***	0.57***	0.22	0.17***	0.15**	-0.01
	(3.86)	(3.53)	(1.07)	(7.25)	(5.08)	(0.94)	(3.69)	(2.10)	(-0.04)
Sells-Buys	-0.95	-0.48	-1.46	-1.29	-0.52	-0.80	0.17	0.05	0.29
Sells=Buys p	0.00	0.00	0.00	0.00	0.01	0.00	0.03	0.58	0.14
N	146,161	30,542	27,095	146,161	30,542	27,095	146,161	30,542	27,095
R^2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 7: Marketable retail order imbalance bins and weekly returns in excess of the market. Week 0 refers to the focal week when retail imbalance is calculated, -1 and 1 refer to the week before and after, respectively. Dependent variable is return in excess of the equal-weighted market return. Panel A contains all weeks, while Panel B restricts to instances where week 0 contains an earnings announcement. The bottom part of the table reports the gap between the dummies on retail buys and retail sells, and the associated p value. Standard errors clustered on the week level.

	CAR[0, 60]			ECAR[0, 60]		
	All	Low	High	All	Low	High
Retail Sells	-0.46***	-1.77***	1.06**	-0.35**	-0.67***	0.12
	(-2.88)	(-6.20)	(2.33)	(-2.50)	(-5.14)	(0.42)
2	-0.63***	-1.06***	1.10	-0.23**	-0.54***	0.64*
	(-4.97)	(-5.73)	(1.58)	(-1.96)	(-5.38)	(1.74)
3	-0.65***	-1.29***	0.77	-0.11	-0.34***	0.96*
	(-5.49)	(-7.70)	(1.05)	(-0.98)	(-3.12)	(1.80)
4	-0.15	-1.04***	2.26***	0.05	-0.26**	1.32***
	(-1.13)	(-5.12)	(3.15)	(0.40)	(-2.35)	(3.38)
Retail Buys	1.15***	-0.78***	3.34***	0.57***	-0.05	1.49***
	(7.19)	(-2.65)	(7.04)	(2.64)	(-0.33)	(2.92)
Sells-Buys	1.61	1.00	2.28	0.92	0.62	1.37
Sells=Buys p	0.00	0.01	0.00	0.00	0.00	0.01
N	140,437	31,860	18,274	140,437	31,860	18,274
R^2	0.00	0.01	0.01	0.00	0.01	0.00

Table 8: Return predictability decomposition. Firm-announcement level regressions of cumulative abonrmal return (CAR) or expected cumulative abnormal return (ECAR) on dummies representing quintiles of retail order imbalance formed over the last ten trading days prior to the announcement. ECAR estimated using the contemporaneous relationship between returns and order imbalance in non-announcing firms. Day 0 refers to the first day on which the earnings news was tradeable. Columns Low and High restrict the sample to the first and fifth quintile of retail trading intensity. Standard errors clustered by trading day. The bottom of the table reports the difference between the retail buys and retail sells and the associated p value.

	Demeaned Effective Spread (basis points)					
	(-5,-1)	(-3,-1)	-1	0	(0,2)	(0,4)
Low Retail	-0.402**	-0.359	-0.124	-0.745**	-0.341*	-0.126
	(0.184)	(0.239)	(0.308)	(0.317)	(0.203)	(0.216)
2	-0.132	-0.0814	-0.098	0.0557	0.105	0.0814
	(0.135)	(0.163)	(0.238)	(0.263)	(0.167)	(0.147)
4	0.666**	0.680**	0.782*	0.747*	0.641**	0.576**
	(0.258)	(0.288)	(0.392)	(0.447)	(0.283)	(0.248)
High Retail	1.538***	1.431**	0.861	2.927***	2.390***	2.164***
	(0.502)	(0.609)	(0.844)	(0.944)	(0.679)	(0.529)
Observations	137,141	137,141	137,141	137,141	137,141	137,141
R-squared	0.107	0.105	0.099	0.108	0.118	0.124
Time FE	YES	YES	YES	YES	YES	YES
Controls	YES	YES	YES	YES	YES	YES
Firm FE	YES	YES	YES	YES	YES	YES

Table 9: Retail activity and demeaned trading costs around earnings announcements. Left-hand-side variables are average demeaned effective spread computed over various windows around earnings announcements. Demeaned effective spread is effective bid-ask spread minus average effective spread over the calendar month before the earnings announcement. Control variables include nominal price, returns from month $t-12$ to $t-2$, time since listing, market capitalization, book-to-market, gross profit margin, book long-term leverage, MAX (lottery demand) and month $t-1$ returns. Additional controls include betas on the three Fama-French factors as well as the momentum factor, idiosyncratic volatility and total volatility, all computed over the past 12 months. Standard errors are double clustered at the permno and month level.

	(-1, 0)			(-3, 2)		
	All	Pre	Post	All	Pre	Post
Size 1 x Q1	-0.23	-0.25**	0.01	-0.52	-0.41**	-0.13
	(-1.07)	(-3.14)	(0.06)	(-1.78)	(-3.05)	(-0.56)
Size 1 x Q5	-1.28***	0.11	-1.36***	-1.77***	0.52**	-2.19***
	(-4.77)	(0.98)	(-6.03)	(-4.64)	(2.74)	(-7.03)
Size 1	0.38*	0.33***	0.07	0.82***	0.50***	0.31
	(2.54)	(6.13)	(0.52)	(3.79)	(5.45)	(1.73)
Size 2 x Q1	0.21	-0.03	0.23	0.23	-0.03	0.26
	(1.21)	(-0.50)	(1.48)	(1.08)	(-0.31)	(1.48)
Size 2 x Q5	-0.85***	0.07	-0.92***	-0.72*	0.23	-0.95***
	(-3.90)	(0.94)	(-4.70)	(-2.35)	(1.39)	(-3.87)
Size 2	0.36**	0.08	0.30**	0.39**	0.10	0.31*
	(3.16)	(1.69)	(2.86)	(2.79)	(1.24)	(2.43)
Size 3 x Q1	0.04	0.01	0.04	-0.01	-0.05	0.05
	(0.26)	(0.09)	(0.26)	(-0.06)	(-0.57)	(0.32)
Size 3 x Q5	-0.78***	0.10	-0.87***	-0.67**	0.18	-0.82***
	(-4.11)	(1.19)	(-4.81)	(-2.77)	(1.31)	(-3.79)
Size 3	0.43***	0.02	0.41***	0.60***	0.09	0.51***
	(4.04)	(0.53)	(4.03)	(4.72)	(1.31)	(4.33)
Size 4 x Q1	0.03	0.04	-0.02	0.23	0.10	0.11
	(0.22)	(1.09)	(-0.13)	(1.38)	(1.45)	(0.78)
Size 4 x Q5	-0.57**	0.07	-0.63***	-0.50*	0.12	-0.62***
	(-3.28)	(1.18)	(-4.03)	(-2.37)	(1.00)	(-3.48)
Size 4	0.34***	0.04	0.30***	0.42***	0.14**	0.30**
	(3.74)	(1.71)	(3.53)	(4.08)	(2.72)	(3.21)
Size 5 x Q1	0.04	-0.01	0.06	0.20	0.08	0.13
	(0.39)	(-0.32)	(0.55)	(1.62)	(1.28)	(1.17)
Size 5 x Q5	-0.22	0.06	-0.27**	-0.19	0.16	-0.36**
	(-1.75)	(0.95)	(-2.73)	(-1.15)	(1.87)	(-2.63)
Size 5	0.26***	0.13***	0.14*	0.37***	0.20***	0.18*
	(3.65)	(3.67)	(2.22)	(4.68)	(5.32)	(2.55)
N	30,355	30,355	30,355	30,355	30,355	30,355

Table 10: Cumulative returns around earnings announcements. Earnings announcements in 2007-2021. Column headers refer to first and last days in return window. 0 is the first trading day on which announcement information is tradeable. Q denotes indicator variables for retail intensity quintiles; Size denotes indicator variables for size quintiles, both sorted in the month prior to the announcement. Pre refers to trading days prior to announcement, post refers to trading days after announcement. Monthly fixed effects. Standard errors clustered by firm and month.

9 References

References

Aase, K. K., Bjuland, T., and Oksendal, B. (2011). Insider trading with partially informed traders. *NHH Dept. of Finance & Management Science Discussion Paper*, (2011/21).

Appel, I. R., Gormley, T. A., and Keim, D. B. (2016). Passive investors, not passive owners. *Journal of Financial Economics*, 121(1):111–141.

Balasubramaniam, V., Campbell, J. Y., Ramadorai, T., and Ranish, B. (2023). Who owns what? a factor model for direct stockholding. *The Journal of Finance*, 78(3):1545–1591.

Bali, T. G., Brown, S. J., Murray, S., and Tang, Y. (2017). A lottery-demand-based explanation of the beta anomaly. *Journal of Financial and Quantitative Analysis*, 52(6):2369–2397.

Banerjee, S., Davis, J., and Gondhi, N. (2021). Choosing to disagree: Endogenous dismissiveness and overconfidence in financial markets. Working Paper.

Barber, B. M., Huang, X., Jorion, P., Odean, T., and Schwarz, C. (2023). A (sub) penny for your thoughts: Tracking retail investor activity in taq. *The Journal of Finance*.

Barber, B. M. and Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *The Journal of Finance*, 55(2):773–806.

Barber, B. M. and Odean, T. (2002). Online investors: do the slow die first? *The Review of Financial Studies*, 15(2):455–488.

Barrot, J.-N., Kaniel, R., and Sraer, D. (2016). Are retail traders compensated for providing liquidity? *Journal of Financial Economics*, 120(1):146–168.

Battalio, R. H., Jennings, R. H., Saglam, M., and Wu, J. (2023). Identifying market maker trades as "retail' from TAQ: No shortage of false negatives and false positives. Working Paper.

Beaver, W. H. (1968). The information content of annual earnings announcements. *Journal of Accounting Research*, pages 67–92.

Beber, A., Brandt, M. W., Cen, J., and Kavajecz, K. A. (2021). Mutual fund performance: Using bespoke benchmarks to disentangle mandates, constraints and skill. *Journal of Empirical Finance*, 60:74–93.

Ben-David, I., Kim, B., Moussawi, H., and Roulstone, D. (2023). Corporate transactions in hard-to-value stocks. *The Review of Corporate Finance Studies*, 12(3):539–580.

Benartzi, S. (2001). Excessive extrapolation and the allocation of 401 (k) accounts to company stock. *The Journal of Finance*, 56(5):1747–1764.

Black, F. (1986). Noise. *The journal of finance*, 41(3):528–543.

Boehmer, E., Jones, C. M., Zhang, X., and Zhang, X. (2021). Tracking retail investor activity. *The Journal of Finance*, 76(5):2249–2305.

Bogousslavsky, V. and Muravyev, D. (2024). An anatomy of retail option trading. Working Paper.

Bryzgalova, S., Pavlova, A., and Sikorskaya, T. (2023). Retail trading in options and the rise of the big three wholesalers. *The Journal of Finance*, 78(6):3465–3514.

Campbell, J. Y., Lettau, M., Malkiel, B. G., and Xu, Y. (2001). Have individual stocks become more volatile? an empirical exploration of idiosyncratic risk. *The Journal of Finance*, 56(1):1–43.

Cohen, L. and Lou, D. (2012). Complicated firms. *Journal of Financial Economics*, 104(2):383–400.

Cohen, L., Malloy, C., and Nguyen, Q. (2020). Lazy prices. *The Journal of Finance*, 75(3):1371–1415.

Curcuru, S., Heaton, J., Lucas, D., and Moore, D. (2010). Heterogeneity and portfolio choice: Theory and evidence. In *Handbook of financial econometrics: Tools and techniques*, pages 337–382. Elsevier.

De Long, J. B., Shleifer, A., Summers, L. H., and Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of Political Economy*, 98(4):703–738.

Décaire, P. H., Sosyura, D., and Wittry, M. D. (2023). Model input choice and valuation disagreement. Working Paper.

DellaVigna, S. and Pollet, J. M. (2009). Investor inattention and friday earnings announcements. *The Journal of Finance*, 64(2):709–749.

Dhar, R. and Zhu, N. (2006). Up close and personal: Investor sophistication and the disposition effect. *Management Science*, 52(5):726–740.

Di Maggio, M., Egan, M., and Franzoni, F. (2022). The value of intermediation in the stock market. *Journal of Financial Economics*, 145(2):208–233.

Di Maggio, M., Franzoni, F., Kermani, A., and Sommavilla, C. (2019). The relevance of broker networks for information diffusion in the stock market. *Journal of Financial Economics*, 134(2):419–446.

Di Maggio, M., Franzoni, F. A., Kogan, S., and Xing, R. (2021). Avoiding volatility: Institutional trading before scheduled news. working Paper.

Edmans, A., Fang, V. W., and Zur, E. (2013). The effect of liquidity on governance. *The Review of Financial Studies*, 26(6):1443–1482.

Engelberg, J. (2008). Costly information processing: Evidence from earnings announcements. In *AFA 2009 San Francisco meetings paper*.

Frazzini, A. and Lamont, O. A. (2007). The earnings announcement premium and trading volume. Working Paper.

Gabaix, X., Koijen, R. S. J., Mainardi, F., Oh, S., and Yogo, M. (2022). Asset demand of u.s. households. Working Paper.

Glosten, L., Nallareddy, S., and Zou, Y. (2021). Etf activity and informational efficiency of underlying securities. *Management Science*, 67(1):22–47.

Golubov, A. and Konstantinidi, T. (2021). Valuation uncertainty. Working Paper.

Gormsen, N. J. and Lazarus, E. (2021). Duration-driven returns. *Available at SSRN 3359027*.

Gormsen, N. J. and Lazarus, E. (2023). Duration-driven returns. *The Journal of Finance*, 78(3):1393–1447.

Greenwood, R., Laarits, T., and Wurgler, J. (2023). Stock market stimulus. *Review of Financial Studies*.

Greenwood, R. and Sammon, M. C. (2022). The disappearing index effect. Working Paper.

Haddad, V., Huebner, P., and Loualiche, E. (2021). How competitive is the stock market? theory, evidence from portfolios, and implications for the rise of passive investing. Working Paper.

Hartzmark, S. M. and Shue, K. (2018). A tough act to follow: Contrast effects in financial markets. *The Journal of Finance*, 73(4):1567–1613.

Hirshleifer, D., Lim, S. S., and Teoh, S. H. (2009). Driven to distraction: Extraneous events and underreaction to earnings news. *The Journal of Finance*, 64(5):2289–2325.

Holden, C. W. and Jacobsen, S. (2014). Liquidity measurement problems in fast, competitive markets: Expensive and cheap solutions. *The Journal of Finance*, 69(4):1747–1785.

Hong, H. and Stein, J. C. (2007). Disagreement and the stock market. *Journal of Economic Perspectives*, 21(2):109–128.

Huberman, G. (2001). Familiarity breeds investment. *The Review of Financial Studies*, 14(3):659–680.

Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1):65–91.

Johnson, T. L. and So, E. C. (2018). Asymmetric trading costs prior to earnings announcements: Implications for price discovery and returns. *Journal of Accounting Research*, 56(1):217–263.

Kacperczyk, M., Van Nieuwerburgh, S., and Veldkamp, L. (2016). A rational theory of mutual funds' attention allocation. *Econometrica*, 84(2):571–626.

Kaniel, R., Liu, S., Saar, G., and Titman, S. (2012). Individual investor trading and return patterns around earnings announcements. *The Journal of Finance*, 67(2):639–680.

Kaniel, R., Saar, G., and Titman, S. (2008). Individual investor trading and stock returns. *The Journal of Finance*, 63(1):273–310.

Kogan, L., Papanikolaou, D., Seru, A., and Stoffman, N. (2017). Technological innovation, resource allocation, and growth. *The Quarterly Journal of Economics*, 132(2):665–712.

Koijen, R. S., Richmond, R. J., and Yogo, M. (2020). Which investors matter for equity valuations and expected returns? Working Paper.

Koijen, R. S. and Yogo, M. (2019). A demand system approach to asset pricing. *Journal of Political*

Economy, 127(4):1475–1515.

Kothari, S. P. and Sloan, R. G. (1992). Information in prices about future earnings: Implications for earnings response coefficients. *Journal of Accounting and Economics*, 15(2-3):143–171.

Kumar, A. (2009). Hard-to-value stocks, behavioral biases, and informed trading. *Journal of Financial and Quantitative Analysis*, 44(6):1375–1401.

Kumar, A. and Lee, C. M. (2006). Retail investor sentiment and return comovements. *The Journal of Finance*, 61(5):2451–2486.

Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, pages 1315–1335.

Lee, C. M. and Ready, M. J. (1991). Inferring trade direction from intraday data. *The Journal of Finance*, 46(2):733–746.

Lev, B. and Gu, F. (2016). *The end of accounting and the path forward for investors and managers*. John Wiley & Sons.

Li, J. and Lin, Z. (2023). Price multipliers are larger at more aggregate levels. Technical report. Working Paper.

Liu, H., Peng, C., Xiong, W. A., and Xiong, W. (2020). Resolving the excessive trading puzzle: An integrated approach based on surveys and transactions.

Livnat, J. and Mendenhall, R. R. (2006). Comparing the post–earnings announcement drift for surprises calculated from analyst and time series forecasts. *Journal of Accounting Research*, 44(1):177–205.

Luo, C., Ravina, E., Sammon, M., and Viceira, L. M. (2021). Retail investors' contrarian behavior around news and the momentum effect. Working Paper.

Ma, L., Tang, Y., and Gomez, J.-P. (2019). Portfolio manager compensation in the us mutual fund industry. *The Journal of Finance*, 74(2):587–638.

Martineau, C. and Mondria, J. (2022). News selection and asset pricing implications. Working Paper.

Martineau, C. and Zoican, M. (2019). Retail trading and analyst coverage. Working Paper.

McLean, R. D., Pontiff, J., and Reilly, C. (2020). Taking sides on return predictability. Working Paper.

Odean, T. (1998). Are investors reluctant to realize their losses? *The Journal of Finance*, 53(5):1775–1798.

Peters, R. H. and Taylor, L. A. (2017). Intangible capital and the investment-q relation. *Journal of Financial Economics*, 123(2):251–272.

Rochet, J.-C. and Vila, J.-L. (1994). Insider trading without normality. *The review of economic studies*, 61(1):131–152.

Savor, P. and Wilson, M. (2016). Earnings announcements and systematic risk. *The Journal of*

Finance, 71(1):83–138.

Seasholes, M. S. and Zhu, N. (2010). Individual investors and local bias. *The Journal of Finance*, 65(5):1987–2010.

Stambaugh, R. F. and Yuan, Y. (2017). Mispricing factors. *The Review of Financial Studies*, 30(4):1270–1315.

Vaarmets, T., Liivamägi, K., and Talpsepp, T. (2019). How does learning and education help to overcome the disposition effect? *Review of Finance*, 23(4):801–830.

van der Beck, P. (2022). Identifying elasticities in demand-based asset pricing. Working Paper.

Welch, I. (2022). The wisdom of the robinhood crowd. *The Journal of Finance*, 77(3):1489–1527.

A Online Appendix

A.1 Trade-off between signal precision and noise trading intensity in Kyle (1985)

In this section, we illustrate the trade-off between signal precision and noise trading intensity in the context of a 2-period Kyle (1985)-style model.[20] We would like to be clear that the model only has one asset, and our comparative statics are comparing investor outcomes *across equilibria*. That being said, our preferred interpretation is that, in practice, there are gains to specialization. So, one could view the stock market as many versions of the one-asset model running in parallel, and based on differences in nose trading intensity and difficulty to value, informed investors pick a single stock to specialize in.

Another important assumption in this model is that retail investors are pure noise traders, in the sense that their order flow is uncorrelated with securities' terminal payoffs. It may be, however, that retail investors have information about future fundamentals(Kaniel et al. (2008), Barrot et al. (2016), Boehmer et al. (2021)). We believe this baseline assumption of retail investors as classical noise traders is conservative, in the sense that it would not naturally discourage informed investors from learning about stocks with a large retail trader presence. As shown in Aase et al. (2011), if noise traders (i.e., retail investors in our setting) have order flow which is correlated with fundamentals, the insider's informational advantage would be relatively smaller, making informed investors less inclined to learn about high retail stocks. In other words, we would naturally bias informed investors from learning about high retail stocks if we modeled retail investors as having information.

A final important assumption in the model is that the market maker cannot cannot separately observe noise trader order flow, but instead only observes total order flow (i.e., the sum of noise trader and informed investor order flow). This seems somewhat inconsistent with our empirical setting, as either (1) a market maker could use the Boehmer et al. (2021) algorithm in real time to identify trades which appear to be retail-initiated and/or (2) the market maker could be directly internalizing retail orders on behalf of a retail brokerage like Robinhood, and thus directly observe some part of retail order flow. As implied by Black (1986), however, a model where the market maker directly observes all noise trader activity will feature no trade. The closest model to capturing this idea is Rochet and Vila (1994), who develop a version of the Kyle (1985) model where the *informed investor* – rather than the market maker – directly observes noise trading activity before submitting their demand. Importantly, in our setting, a market maker either using the Boehmer et al. (2021) algorithm or observing internalized retail order flow would still not be able to *completely* isolate noise trader activity, and thus the model's main predictions should still apply to our empirical exercises.

[20] This section borrows heavily from Alex Chinco's "Two Period Kyle (1985) Model" notes.

A.1.1 Model setup

The model has two trading periods, $t = 1$ and $t = 2$. There is a single risky asset whose value is distributed:

$$v \sim N(0, \sigma_v^2) \tag{A1}$$

There is a *strategic* risk-neutral informed investor who receives an unbiased signal before the first trading period:

$$s = v + \epsilon \tag{A2}$$

where v is the true value of the asset and ϵ is signal noise. ϵ is independent of v and normally distributed with mean zero and standard deviation σ_ϵ. This implies that $s \sim N(v, \sigma_\epsilon^2)$.

The informed investor submits demands to a set of *competitive* risk-neutral market makers at times 1 and 2, y_1 and y_2. To prevent prices from being fully revealing, there are a group of noise traders who submit random demands z_1 and z_2, where the z_t are independent and normally distributed with mean zero and standard deviation σ_z.

The set of competitive market makers observe total order flow x_t each period:

$$x_t = y_t + z_t \tag{A3}$$

There is perfect competition among market makers, so they must set prices equal to the expected fundamental value of the asset given total demand:

$$p_1 = E[v|x_1] \quad \text{and} \quad p_2 = E[v|x_1, x_2] \tag{A4}$$

In period 1, the informed investor chooses demand y_t to solve:

$$H_0 = max_{y_1} E\left[(v - p_1)\, y_1 + H_1 | s\right] \tag{A5}$$

where H_{t-1} is the informed investor's value function entering period t.

In period 2, they choose y_2 to maximize:

$$H_1 = max_{y_2} E\left[(v - p_2)\, y_2 | s, p_1\right] \tag{A6}$$

An equilibrium is made up of two components: (1) a linear demand rule for the informed investor in each period:

$$y_t = \alpha_{t-1} + \beta_{t-1} s \tag{A7}$$

And (2) a liner pricing rule for the market makers in each period:

$$p_t = \kappa_{t-1} + \lambda_{t-1} x_t \tag{A8}$$

The informed investor updates their beliefs about v after observing s. Their posterior beliefs about the mean and variance are:

$$\mu_{v|s} = \left(\frac{\sigma_v^2}{\sigma_v^2 + \sigma_\epsilon^2}\right) \times s \quad \text{and} \quad \sigma_{v|s}^2 = \left(\frac{\sigma_\epsilon^2}{\sigma_v^2 + \sigma_\epsilon^2}\right) \times \sigma_v^2 \tag{A9}$$

where going forward, we will use θ in place of $\left(\frac{\sigma_v^2}{\sigma_v^2+\sigma_\epsilon^2}\right)$.

The market makers extract an unbiased signal about v from total demand. Substituting in the informed trader's demand rule, the $t = 1$ signal is:

$$v = \frac{x_1}{\beta_0} - \epsilon - \frac{z_1}{\beta_0} \tag{A10}$$

A.1.2 Solving the model

Given the market makers' zero profit condition, $\kappa_0 = 0$ and

$$\kappa_1 = E[v|x_1] - \lambda_1 E[x_2|x_1] = p_1 - \left(\theta\mu_{s|x_1} - p_1\right) = p_1 \tag{A11}$$

where the last equality comes from $\theta\mu_{s|x_1} = p_1$.

Substituting in the market makers' linear pricing rule into H_1

$$H_1 = max_{y_2} E\left[\left(v - \kappa_1 - \lambda_1 x_2\right) y_2 | s, p_1\right] \tag{A12}$$

Taking the first order condition with respect to y_2 yields optimal demand:

$$y_2 = -\frac{p_1}{2\lambda_1} + \frac{\theta}{2\lambda_1} s \tag{A13}$$

so $\alpha_1 = -\frac{p_1}{2\lambda_1}$ and $\beta_1 = \frac{\theta}{2\lambda_1}$.

With this, we can partially solve for the market makers' price impact coefficient, λ_t, in period 2:

$$\lambda_1 = \frac{Cov[x_2, v|x_1]}{Var[x_2|x_1]} = \frac{\beta_1 \sigma_{v|x_1}^2}{\beta_1^2 \sigma_{s|x_1}^2 + \sigma_z^2} \tag{A14}$$

Now, turning to the period one solution, we start by taking a guess at at the informed investors' value function which we will verify later:

$$E[H_1|s] = \phi_1 + \omega_1 \left(\mu_{v|s} - p_1\right)^2 \tag{A15}$$

Substituting in the price impact and demand coefficients into H_0 yields:

$$H_0 = max_{y_1} E\left[\left(v - p_1\right) y_1 + \phi_1 + \omega_1 \left(\theta s - p_1\right)^2 | s\right] \tag{A16}$$

Taking the first order condition with respect to y_1 implies:

$$y_1 = \frac{\theta}{2\lambda_0}\left(\frac{1-2\omega_1\lambda_0}{1-\omega_1\lambda_0}\right)s \tag{A17}$$

With all this, we can now solve for the time 1 price impact coefficient:

$$\lambda_0 = \frac{Cov[x_1, v]}{Var[x_1]} = \frac{\beta_0\sigma_v^2}{\beta_0^2\sigma_s^2+\sigma_z^2} \tag{A18}$$

To verify the guess about H_1, substitute the equilibrium coefficients for demands and prices into Equation A15:

$$H_1 = \left[\frac{1}{2\lambda_1}\left([v-\theta s]+\frac{1}{2}[\theta s - p_1]-\lambda_1 z_2\right)(\theta s - p - 1)\,|s\right] \tag{A19}$$

which simplifies to:

$$H_1 = \text{Constant} + \frac{1}{4\lambda_1}\left(\mu_{v|s}-p_1\right)^2 \tag{A20}$$

This reveals that $\omega_1 = \frac{1}{4\lambda_1}$ and that H_1 is consistent with the original guess.

To solve the model, start with some initial guess for $\hat{\beta}_0$, and use this to compute the other equilibrium coefficients. This can be done in stages, first computing $\hat{\sigma^2_{v|x_1}}$ and $\hat{\sigma^2_{s|x_1}}$, and then using these to compute $\hat{\lambda_1}$:

$$\begin{aligned}
\hat{\lambda_0} &= \frac{\hat{\beta}_0\sigma_v^2}{\hat{\beta}_0^2\sigma_s^2+\sigma_z^2}\\
\hat{\sigma^2_{v|x_1}} &= \frac{\hat{\beta}_0^2\sigma_\epsilon^2+\sigma_z^2}{\hat{\beta}_0^2\sigma_s^2+\sigma_z^2}\sigma_v^2\\
\hat{\sigma^2_{s|x_1}} &= \frac{\sigma_z^2}{\hat{\beta}_0^2\sigma_s^2+\sigma_z^2}\sigma_s^2\\
\hat{\lambda_1} &= \frac{1}{\sigma_z}\sqrt{\frac{\theta}{2}\left(\hat{\sigma^2_{v|x_1}}-\frac{\theta}{2}\hat{\sigma^2_{s|x_1}}\right)}
\end{aligned} \tag{A21}$$

A solution has been found when the distance between the guess $\hat{\beta}_0$ and $\frac{\theta}{2\hat{\lambda_0}}\left(\frac{1-2\hat{\omega_1}\hat{\lambda_0}}{1-\hat{\omega_1}\hat{\lambda_0}}\right)$ has been minimized which is a condition $\hat{\beta}_0$ has to satisfy in equilibrium.

A.1.3 Simulation Results

For each set of parameters, we simulate the economy 10,000 times and compute averages of the insider's total profit, defined as $x_1 \times (v-p_1) + x_2 \times (v-p_2)$. Figure A1 plots the insider's profit against the imprecision of their signal (σ_ϵ) for several values of noise trading intensity σ_z. In all simulations, we set fundamental volatility, σ_v, to one.

Unsurprisingly, the insider's profit is monotonically decreasing in signal imprecision (i.e., moving from left to right), and is monotonically increasing in noise trader intensity. The more interesting result is that the insider's profit can be lower in a high noise trading intensity stock (e.g., the stock represented by the yellow line) than a low noise trading intensity stock (e.g., the stock represented by the blue line) if the precision of their signal is sufficiently higher in the low noise trading intensity stock.

Mapping this back to our empirical results, our preferred interpretation of a hard to value stock is one that, for a given amount of learning energy expended, investors receive a relatively less precise signal. Specifically, consider the multi-asset noisy rational expectations equilibrium with endogenous learning of Kacperczyk et al. (2016). Suppose that an investor i can allocate attention K_{ij} to asset j to receive a signal with variance $1/\left(c_j K_{ij}\right)$ i.e., with a precision of $c_j K_{ij}$ (the model features a transformation which makes the assets uncorrelated). Further, suppose that each investor i has an overall attention budget of K, so $\sum_{j=1}^{J} K_{ij} < K$.

In this setting, c_j is an asset j-specific parameter that governs how easy it is to learn about that asset. Lower values of c_j imply that for a given amount of learning (K_{ij}), investors will receive a less precise signal i.e., lower values of c_j imply that a stock is harder to value. So, mapping this back to the example in Figure A1, suppose that the red line represents a low retail stock ($\sigma_z = 1$) while the yellow line represents a high retail stock ($\sigma_z = 2$). Suppose that the total attention constraint $K = 4$, $c_{red} = 1$ and $c_{yellow} = 0.1$. If an investor allocates all their attention to the low retail stock, their signal will have variance 1/4 (standard deviation 0.5), while if the investor allocates all their attention to the high retail stock, their signal will have variance 2.5 (standard deviation of about 1.58). With these parameters, the informed investor would make more money allocating all their attention to the low retail stock than to the high retail stock, even though the high retail stock has double the noise trading intensity.

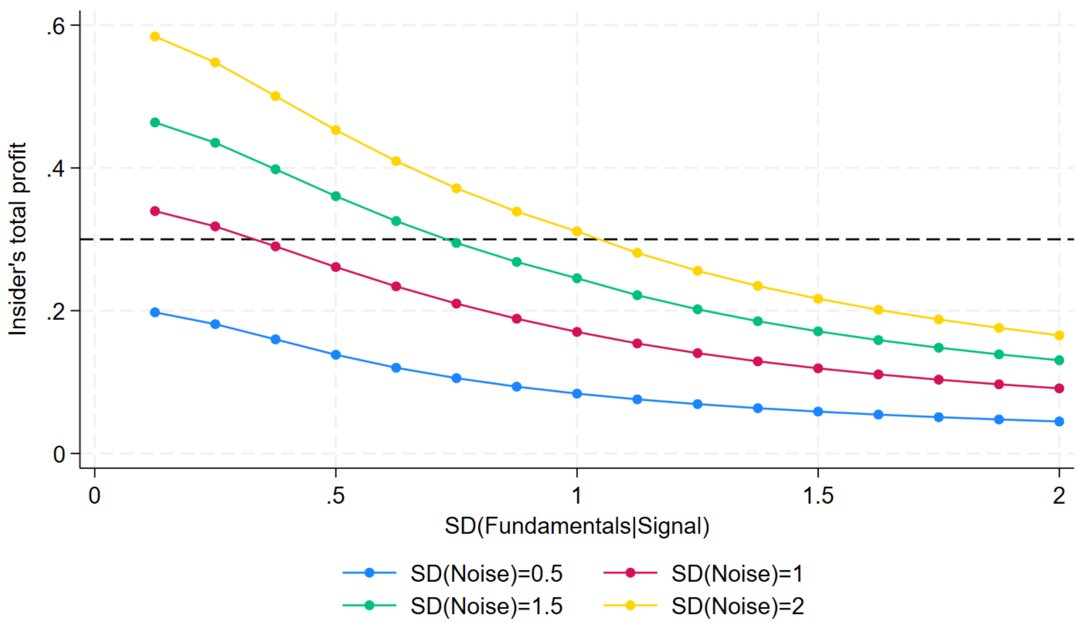

Figure A1: Insider's profits as a function of noise trading intensity and signal precision. Each point represents the average of the insider's total profit in periods 1 and 2 across 10,000 simulations. Fundamental volatility (σ_v) is fixed at 1.

A.2 BJZZ Algorithm Validation

While the Boehmer et al. (2021) (BJZZ) algorithm has gained significant popularity since its publication, several papers have questioned its accuracy. For instance, Barber et al. (2023) argue that the algorithm results in a high number of false negatives, with only 35% of their sample of retail trades being identified as retail, and among those, only 72% being correctly signed. Similarly, Battalio et al. (2023) find that less than 28% of their sample of retail trades are correctly identified by the BJZZ algorithm. Among these identified trades, however, 94% are correctly signed. Battalio et al. (2023) also find evidence of false positives, particularly when institutional investors use Alternative Trading Systems (ATSs), Single Dealer Platforms (SDPs), or wholesalers. In these cases, the BJZZ algorithm can mistakenly identify these institutional trades as retail between 33% and 78% of the time, depending on the platform used.

The natural question, then, is how the presence of these false positives and false negatives affects our results. A potential concern for our applications is that the noise in the BJZZ algorithm is correlated with a firm being hard-to-value. One might be worried that the BJZZ algorithm has more false positives and/or fewer false negatives in hard-to-value stocks. This would lead hard to value stocks to mechanically have more "retail" trades, as identified by BJZZ.

In this section we provide detailed calculations that suggest at most a minor degree of bias in our estimates. Firstly, recall that for most of our exercises, we are only concerned with the ranking of stocks based on gross retail activity, and any "level" effect in false positives or false negatives need not matter for the ranking at all. Secondly, in many of our applications we are interested only in the share of retail trades, rather than the direction. In cases where the direction of retail trades matters, we document that the results are robust to using the improved signing algorithm advocated by Barber et al. (2023) that is based on the Lee and Ready (1991) method. Thirdly, and most importantly, we can leverage the cross-sectional estimates of both Battalio et al. (2023) and Barber et al. (2023) to quantify the degree to which BJZZ misstates retail trading intensity.

A.2.1 Accounting for False Negatives

Table 4 of Battalio et al. (2023) shows that certain stock-level characteristics predict the success rate of the BJZZ algorithm: penny spread, trade price, stock volatility and market cap are all negatively related to the algorithm success rate. The R-squared from their regression of the BJZZ algorithm's success rate on all five of these characteristics together is 0.02. This indicates that, while each characteristic is a statistically significant predictor of the algorithm's accuracy, collectively, the characteristics do not explain much of the total variation in the presence of false negatives. Already, this observation suggests that for our main empirical exercises ranking stocks on past retail activity, mis-classifications by the BJZZ algorithm are unlikely to bias our results. The low total explanatory power of the characteristics suggests that the noise in the BJZZ algorithm looks more like classical measurement error – in the sense that it is not strongly explained by variation in the stock-level characteristics – and therefore it's unlikely that such noise would systematically bias our results.

Let us put the estimates in Table 4 of Battalio et al. (2023) into context. First, although our sample of high retail stocks have higher *percentage* average spreads than low retail stocks (see Table A13), they have similar *dollar* average effective spreads (Holden and Jacobsen, 2014). Specifically, high retail stocks have average effective spreads of 5.69 cents, while low retail stocks have average effective spreads of 6.49 cents (i.e., an average difference of 8/10ths of a cent). In addition, high retail stocks have an average daily effective spread of a penny (or less) on 12.55% of trading days, while low retail stocks have a penny spread (or less) on 5.13% of trading days. Together, these facts suggest that our collection of high retail stocks would have more false negatives (i.e., a lower success rate) than low retail stocks on average. Further, our high retail stocks have relatively larger price volatility than low retail stocks – again implying a larger rate of false negative classifications on average. On the other hand, our high retail stocks have low nominal prices and lower market capitalization. So, on these dimensions, we would expect a smaller number of false negatives in our pool of high retail stocks.

Given that these forces push in opposite directions, the logical next step is to quantify the relative magnitudes of these effects. In a series of univariate regressions of the BJZZ algorithm's success rate on each of the four characteristics mentioned above, Battalio et al. (2023) find a coefficient of -0.03 on penny spread (a dummy variable equal to 1 if all retail trades are executed when the width of the NBBO was a penny), -0.003 on log(avg. retail trade price), -0.13 on stock volatility (measured using the standard deviation of daily returns) and -0.004 on log(market capitalization).[21] As a baseline suppose that 20% of all trades are truly retail-initiated, while 80% are initiated by institutions. This baseline is based on the idea that we observe an average retail share of 6.37% in Table 1, and Battalio et al. (2023) argue that 28% of retail trades are correctly identified – so ignoring false positives (which we will address separately below), and grossing this up (0.0637/0.28) yields a number close to 20%. Also, in this baseline we will assume that our ranking is pure noise i.e., there is no difference in the "true" retail share of trading volume between our high and low retail portfolios. With this baseline in mind, we can quantify the expected spread in the observed share of retail trading volume based on each of these individual characteristics.

For example, we find that our sample of low retail stocks have a standard deviation of daily returns of 0.022, while high retail stocks have a standard deviation of 0.041. The intercept in Battalio et al. (2023)'s regression is 0.37, so for low retail stocks we would expect a success rate of $0.37 + 0.022 \times -0.13 = 0.3671$, while for high retail stocks we would expect a success rate of $0.37 + 0.041 \times -0.13 = 0.3647$. Again, assuming that there are no false positives (which we will address separately below), for our low retail, of the 20% of true retail trades, 36.71% would be classified by BJZZ as retail, for an overall retail share of 7.34%. A similar calculation implies that for our high retail stocks, the expected retail share would be 7.29%. So, only focusing on unconditional differences in volatility, our sample of high retail stocks are expected to have a 5 basis point *lower* retail share of trading volume than low retail stocks. This magnitude is economically tiny, as in

[21] We cannot directly observe the fifth characteristic examined in Battalio et al. (2023), log(avg. retail trade size), so we omit it from this analysis.

Table 1, our high retail stocks have a 13.26% higher retail share of trading volume than low retail stocks.

Following this logic, we can perform a similar calculation for each of the other stock-level characteristics. We obtain data on average retail buy and sell prices from the WRDS intraday indicators suite, and find that high retail stocks have an average log trade price of 1.78, while low retail stocks have an average log trade price of 3.51. This would imply a 10 basis point *higher* retail share for our high retail stocks relative to our low retail stocks. In addition, our low retail stocks have an average log market capitalization of 21.33, while our high retail stocks have an average log market capitalization of 19.13. This would imply a 18 basis point *higher* retail share for our high retail stocks. Finally, while we cannot observe exactly the measure of penny spread used by Battalio et al. (2023), our stock-day-level proxy from the WRDS intraday indicators suite would imply a 4 basis point *lower* retail share for our high retail stocks. So, as a rough approximation, if we add together these effects, we would expect a -5bp + 10bp + 18bp - 4bp = 19 basis points higher retail share of trading volume in our sample of high retail stocks. This suggests that differences in false-negative rates driven by stock-level characteristics are unlikely to substantially bias our results because, as mentioned above, 19 basis points is small compared to the unconditional difference in the retail share of trading volume between our sample of high and low retail stocks which is above 13 percentage points.

A.2.2 Accounting for False Positives

False negatives, however, are only part of the story about the effect of classification errors in the BJZZ algorithm on our results, as one may also be worried about the presence of false positives. Table 7 of Battalio et al. (2023) reports that stocks with single-penny spreads, low volatility and large market capitalization have relatively more false positives i.e., institutional trades classified as retail orders by BJZZ. As with the results on the false negative rates, however, the collective R-squared of all these characteristics is small at 0.04, suggesting that stock-level characteristics do not explain a significant share of the variation in the false positive rate.

To quantify the magnitudes of these effects, we will again return to our baseline example where our rankings are pure noise and 20% of trades are truly retail-initiated, so the false positive rates will apply to the 80% of trades which are truly initiated by institutions. In a regression of the BJZZ algorithm's false positive rate on the three characteristics mentioned above (as well as turnover, which is insignificant, and the participation rate, which we cannot observe in our data), Battalio et al. (2023) find a coefficient of 0.04 on penny spread, -0.44 on volatility and 0.01 on log market capitalization. Following the logic of the calculations above for false negatives, differences in volatility would imply that our high retail stocks would have a 67 basis point lower retail share of trading volume relative to our sample of low retail stocks. For log(market capitalization), the implied effect is a 1.76% lower retail share for our high retail stocks. And finally, for penny spread, the implied effect is a 24 basis point higher retail share for our high retail stocks. Collectively, these results imply a -2.19% lower retail share of trading volume for our high retail stocks relative to our low retail stocks. While this effect is economically much larger than the effect of false negatives, the sign is negative, suggesting that differences in false positive rates correlated with firm characteristics would work against our findings.

To summarize, given the results in Battalio et al. (2023) and differences in characteristics between our high and low retail portfolios, we have three main conclusions. First, while stock-level characteristics are statistically strong predictors of variation in the false positive and false negative classification rate for the BJZZ algorithm, they explain a relatively small share of the total variation in these errors. Second, while false negatives would work in the same direction as our results, the effect is economically very small. Finally, the effect of false positives is economically large, but they would likely work against our results. Overall, from this analysis we conclude that errors in the BJZZ algorithm are unlikely to be a major source of bias in our results.

A.2.3 Alternative Estimates of False Negatives

So far, we have focused on the results in Battalio et al. (2023) to understand the effects of possible classification errors by the BJZZ algorithm on our results. As additional robustness, we perform a similar exercise leveraging the results in Barber et al. (2023). Barber et al. (2023) find that in stocks with larger nominal spreads, odd spreads and high nominal prices, the BJZZ algorithm has larger false negative rates.[22] As with the results in Battalio et al. (2023), however, for a large set of 13 stock-level and economy-level characteristics, the R-squared is only 0.018 (including broker and firm fixed effects increases the R-squared to 0.043), suggesting that again, while the characteristics are statistically significant predictors of variation in BJZZ error rates, they explain a small share of the total variation in false negatives.

To quantify the expected effects of differences in characteristics on our results, we start from the baseline example where our rankings are pure noise and 20% of trades are truly retail-initiated. In a regression of the BJZZ algorithm's false negative rate on all the characteristics (but excluding broker fixed effects, as we cannot observe those in our data), Barber et al. (2023) find a coefficient of -0.0731 on dummy variable for a spread between 5 cents and 10 cents, -0.0922 on a dummy variable for a spread greater than 10 cents, and -0.0138 on log(nominal price). The authors do not report a constant term in this regression, so we use their unconditional false negative rate of 65%. All the other characteristics are marginally significant or insignificant except for odd spread – we cannot observe the spread faced at the time the trade was placed. So, following the same logic we used to quantify the effect of false negatives based on the results in Battalio et al. (2023), we will apply these regression estimates to the differences in characteristics between our sample of high and low retail stocks.

[22] Note that there is an apparent tension between the results in Battalio et al. (2023) and Barber et al. (2023). Specifically, Battalio et al. (2023) find that stocks with small (one penny) spreads have lower BJZZ success rates (i.e., more false negatives) while Barber et al. (2023) find more false negatives for stocks with large spreads. As argued in Battalio et al. (2023), a possible source of these seemingly conflicting results is that the trades studied in Barber et al. (2023) may not be representative of the "average" retail trade. This is because Barber et al. (2023) focus on a set of 85,000 trades placed by the authors which averaged less than $100 each, while Battalio et al. (2023) focus on a larger universe of 53 million trades in data directly from wholesalers.

Our sample of high retail stocks have a 54 basis point lower probability than our low retail stocks of having an average daily spread greater than 5 cents, but less than or equal to 10 cents. This would imply a less than 1 basis point difference in the retail share of trading volume between our high and low retail stocks. In addition, our high retail stocks have a 1.45% lower probability than our low retail stocks of having a spread greater than 10 cents. This would imply a 2.7 basis higher retail share of trading volume for our high retail stocks. Finally, as discussed above, our high retail stocks have a log average price of 1.78, while our low retail stocks have a log average price of 3.51. This would imply a 48 basis point higher retail share for our high retail stocks relative to our low retail stocks. So, the collective net effect of these differences is small – implying a roughly 51 basis point higher retail share of trading volume in our high retail stocks – which is small relative to the unconditional spread of 13.26%.

To summarize, from our analysis based on the results Barber et al. (2023) and Battalio et al. (2023), we conclude that it is unlikely that the effect of mis-classifications by the BJZZ algorithm are driving our results.

A.2.4 Other Validation of the BJZZ algorithm

As another way to validate our use of the BJZZ algorithm, we compare our ranking of stocks on retail trading intensity to rankings based on other measures of retail trading activity in Table A1.

In panel A, we form 5 quintiles based on the number of Robinhood users from Robintrack, and compare these to 5 quintiles formed on retail trading intensity, defined as $\text{RSVOL}_{i,t} = \frac{\text{RBuy}_{i,t}+\text{RSell}_{i,t}}{\text{Volume}_{i,t}}$. Although we expect these two quantities to be related, they may not be perfectly correlated, as the number of users holding a stock is not necessarily a measure of trading intensity.[23] We find that almost 70% of stocks in the bottom quintile of retail trading intensity are in the bottom two quintiles of Robintrack users. Similarly, over 60% of stocks in the top quintile of retail trading intensity are in the top two quintiles of Robintrack users.

In panels B and C of Table A1, we form 5 quintiles based on the fraction of total volume coming from internalized orders at Citadel and Virtu, two of the largest wholesalers for retail order flow. Owing to SEC rule 605, wholesalers need to make available on their websites data with statistics on price improvement for their internalized orders. We define wholersaler internalization intensity as total shares from internalized trades at each wholesaler divided by total volume in CRSP. We show that the overlap between these wholesaler-based measures and retail trading intensity calculated using BJZZ is even higher than the overlap with Robintrack activity in panel A. For example, almost 100% of stocks in the top quintile of retail trading intensity are in the top two quintiles of wholesaler internalization intensity for both Citadel and Virtu. Similarly, nearly all stocks in the bottom quintile of retail trading intensity are in the bottom two quintiles of internalization intensity.

Part of this relationship is mechanical. The BJZZ algorithm is designed to identify internalized

[23] As discussed in Luo et al. (2021), in their dataset from a large discount retail brokerage, a small number of day-traders make up the majority of dollar trading volume.

orders that receive price improvement. And wholesalers report orders they *choose* to internalize (there is always the option to send the order directly to the exchange), which may be those that receive price improvement. Further, because the BJZZ algorithm only identifies sub-penny price improvements, the internalized trades with price improvement of more than a penny will not be classified as retail-initiated trades. That being said, at a high level, the results in Table A1 show that ranking stocks on our BJZZ-based measure is consistent with ranking stocks on other measures of retail trading activity.

Panel A: Quintiles of Robintrack Users		Low	2	3	4	High
Quintile of Retail Trading Intensity	Low	36.3%	32.6%	19.3%	9.3%	2.5%
	2	21.8%	26.8%	25.5%	18.4%	7.5%
	3	17.4%	17.5%	21.7%	25.5%	17.9%
	4	14.5%	10.4%	16.6%	22.3%	36.2%
	High	9.8%	12.4%	16.8%	24.5%	36.6%
Panel B: Quintiles of Virtu 605 Trades		Low	2	3	4	High
Quintile of Retail Trading Intensity	Low	69.6%	24.5%	5.2%	0.6%	0.1%
	2	26.2%	49.8%	21.4%	2.5%	0.1%
	3	3.6%	24.6%	54.2%	17.0%	0.5%
	4	0.3%	1.1%	19.2%	65.7%	13.7%
	High	0.1%	0.1%	0.3%	14.4%	85.2%
Panel C: Quintiles of Citadel 605 Trades		Low	2	3	4	High
Quintile of Retail Trading Intensity	Low	75.6%	21.4%	2.5%	0.2%	0.3%
	2	22.0%	57.0%	19.8%	1.1%	0.0%
	3	2.3%	21.0%	61.0%	15.3%	0.3%
	4	0.1%	0.5%	16.9%	70.6%	11.8%
	High	0.1%	0.1%	0.0%	12.9%	86.9%

Table A1: Validation. Each month, we form 5 quintiles on retail traing intensity, defined as $\text{RSVOL}_{i,t} = \frac{\text{RBuy}_{i,t}+\text{RSell}_{i,t}}{\text{Volume}_{i,t}}$. In each panel, we compute the percentage of firms in each retail trading intensity quintile that fall into quintiles formed on other measures of retail investor activity. In panel A, we form quintiles based on the number of Robintrack users. In panels B and C, we compute quintiles based on wholsaler internalization intensity, defined as internalized order volume divided by total CRSP volume, for Virtu and Citadel.

A.3 Variable definitions

This section contains a list of all the variables used in our empirical exercises and their associated definitions.

- **Retail Trading:** the sum of retail-initiated buys and retail-initiated sells (in units of shares bought and sold) divided by total trading volume in CRSP, multiplied by 100.
- **Turnover:** total trading volume in CRSP divided by shares outstanding, multiplied by 100.
- **Retail Turnover:** retail initiated buys plus retail initiated sells (in units of shares bought and sold) divided by shares outstanding, multiplied by 100.
- **Cap:** market capitalization, calculated as price times shares outstanding.
- **Age:** time in months since listing in CRSP.
- **Prc:** nominal price.
- **Past R:** cumulative stock return from month $t-12$ to month $t-2$.
- **B/M:** book to market ratio from the WRDS ratios suite.
- **E/P:** earnings to price ratio from the WRDS ratios suite.
- $\beta_{\mathbf{CAPM}}$**:** beta with respect to the market factor from the WRDS beta suite.
- **100-Inst.:** 100 minus the percentage of the stock's shares owned by institutional investors, defined as all 13F-filing institutions.
- **CF:** cashflow duration from Gormsen and Lazarus (2023). A composite score with high duration firms having high investment, low profitability, low beta, and low payout.
- $K_{\mathbf{Int}}$**:** the sum of capitalized SG&A (selling general and administrative expensive) and R&D (research and development expense) over market capitalization from Peters and Taylor (2017).
- $K_{\mathbf{Know}}$**:** capitalized R&D (research and development expense) over market capitalization from Peters and Taylor (2017).
- $K_{\mathbf{Org}}$**:** capitalized SG&A (selling general and administrative expensive) over market capitalization from Peters and Taylor (2017).
- **PAT:** market value of patents over the past 5 years, divided by market capitalization. Market value of patents is calculated using the methodology in Kogan et al. (2017).
- **VU:** valuation uncertainty from Golubov and Konstantinidi (2021). A score calculated using the distribution of possible valuations based on multiples analysis, with wider distributions implying more valuation uncertainty.

- **Mispricing:** a composite score based on data in Stambaugh and Yuan (2017) who report the average rank of a stock in eleven anomaly portfolios. We transform their data into a measure of difficult-to-value by calculating the distance of the rank to its theoretical mean. High values of Mispricing, therefore, imply that the stock is a member of more extreme anomaly portfolios.

- **SD:** standard deviation of daily returns.

- **Ivol$_t$:** trade-based intraday volatility, computed as the sum of squared (centered) 1-second returns, following Holden and Jacobsen (2014).

- **λ2:** the coefficient from a regression of percentage change in price on the square-root of signed dollar order imbalance (Holden and Jacobsen, 2014). Designed to be a proxy for λ in Kyle (1985), with higher values denoting more price impact for a given order imbalance. For more detail, see the *WRDS Intraday Indicators Formula Note*.

- **Espread:** percentage effective bid-ask spread, defined as the weighted average percent difference between the trade price and the midpoint, where the average is taken across all trades in a given day, and the weights are proportional to the dollar value of each trade.

- **Rspread:** Percentage realized bid-ask spread, defined as the weighted average percent difference between the trade price and the midpoint five minutes after the trade, where the average is taken across all trades in a given day, and the weights are proportional to the dollar value of each trade.

- **Ann. Return:** the return on the first trading day the earnings information could have been traded on during normal market hours. For example, if earnings were released at 2pm on a given trading day, that day's return will be the announcement day return. If earnings were released at 5pm on a trading day, the next trading day's return will be the announcement return.

- **SUE:** standardized unexpected earnings. Actual earnings minus mean expected earnings, divided by the price the day before the earnings announcement date. Both the actual and expected earnings are from the IBES unadjusted summary file.

- **Idiosyn. SUE:** idiosyncratic standardized unexpected earnings. Following Glosten et al. (2021), we estimate a regression of SUE on market-wide value-weighted average SUE and SIC-2 value-weighted average SUE in 5-year rolling windows. The idiosyncratic component of SUE are the residuals from this regression in the last year of the 5-year rolling window.

- **System. SUE:** systematic standardized unexpected earnings. Following Glosten et al. (2021), we estimate a regression of SUE on market-wide value-weighted average SUE and SIC-2 value-weighted average SUE in 5-year rolling windows. The systematic component of SUE is the fitted values from this regression in the last year of the 5-year rolling window.

- **Analysts Disp.:** analyst dispersion, defined as the standard deviation of analyst estimates from the IBES unadjusted summary file.
- Market-adjusted return: Return minus return on the value-weighted market portfolio, following Campbell et al. (2001).
- **mroibvol:** marketable retail order imbalance. The ratio of retail-initiated buys minus retail-initiated sells to retail-initiated buys plus retail-initiated sells (Boehmer et al., 2021).
- **Hard to Value (HTV) score:** the first principal component of turnover, dispersion in analyst forecasts and idiosyncratic volatility following Ben-David et al. (2023).

A.4 Retail favoring versus institutional avoidance

As discussed in the main body of the paper there are multiple reasons why a stock could be classified as having a high retail share of total trading volume. One way is that the stock has high retail-initiated turnover (RTO), defined as retail buys plus retail sells, divided by shares outstanding. For example, in 12/2019, when sorting stocks into 5 portfolio on RTO, the bottom portfolio has an average RTO of 13 basis points, while the top portfolio has an average RTO of 4.14%. To put this last number in perspective, across all stocks in 12/2019, the average total monthly turnover is 18%. And for reasons discussed in the main body of the paper, our estimate of retail-initiated turnover may understate true trading by retail investors due to the the methodology in Boehmer et al. (2021) only identifying internalized retail market orders.

Another way a stock could be classified as high retail, however, is through institutional investor avoidance. Specifically, one could imagine a stock that has relatively low RTO, but also low overall turnover, thus making the retail *share* of trading volume relatively larger. This is because the retail share of trading volume is the ratio of RTO to overall turnover. This raises the concern that a sort on the retail share of trading volume is actually a sort on overall turnover.

To determine whether a stock is high retail because retail investors favor the stock, or institutions avoid the stock, we perform a double sort. First, we sort stocks into quintiles on overall turnover and then, within each of these quintiles, we sort into 5-sub portfolios based on retail-initiated turnover. The idea is that a stock in one of the relatively low RTO portfolios may still have a high retail share of trading volume if it is in a low overall turnover portfolio. Similarly, a stock with relatively high retail-initiated turnover may have a low retail share of trading volume if institutions heavily trade those stocks i.e., overall turnover is relatively high.

Table A2 contains the results. We report overall turnover, defined as total shares traded in a month divided by shares outstanding. By construction, the retail share is increasing in the retail-initiated turnover, the second sorting variable. Note however, that the differences in retail share of trading volume are very similar across the columns, indicating that our retail share of trading volume measure is not just picking up differences in overall turnover. This similarity instead suggests that our sort in the main body of the paper on retail share of trading volume is indeed its own dimension of cross-sectional heterogeneity.

Retail Share of Trading Volume

		Quintile of Turnover (1st sort)				
		1	2	3	4	5
RTO Quintile (2nd Sort)	1	3.81	2.19	2.09	2.18	2.77
	2	4.99	3.10	2.88	3.00	4.33
	3	6.53	4.09	3.71	3.96	6.38
	4	9.39	5.83	5.17	5.62	9.72
	5	16.21	12.58	11.49	12.04	15.47

Table A2: Results of double sort on overall turnover and retail-initiated turnover. The retail share of trading volume is defined as total retail buys plus total retail sells in a month, divided by total trading volume.

A.5 Persistence of Retail Trading

As discussed in the main body of the paper, retail trading activity is persistent. Table A3 shows that stocks in the highest quintile in terms of retail share of trading have a 66% probability of remaining in the top quintile, and an almost 90% probability of remaining in the top two quintiles 12 months in the future.

Panels B and C of Table A3 repeat the same transition-probability analysis, but also condition on the market capitalization of the stock at time $t = -12$. Again we see substantial persistence in portfolio assignments over time. Among small stocks (those in the bottom 20% of market capitalization) with the highest share of retail trading, over 70% are in the top two quintiles 12 months later. Among large stocks (those in the top 20% of market capitalization) this persistence is considerably stronger, a full 90% of stocks in the high retail quintle are in the top two quintiles 12 months later, with over 66% staying in the top bin.

Panel A.

$t=-12$	Retail Portfolio at $t=0$ 1	2	3	4	5
1	53.4	27.6	12.3	4.9	1.7
2	28.5	35.5	23.9	9.7	2.4
3	13.1	25.6	34.2	21.4	5.7
4	5.1	10.7	24.3	40.3	19.6
5	1.7	2.3	6.2	23.9	66.0

Panel B. Small stocks only.

$t=-12$	Retail Portfolio at $t=0$ 1	2	3	4	5
1	37.8	24.8	17.1	11.8	8.5
2	22.9	24.9	21.7	17.5	13.0
3	13.8	19.5	23.2	22.7	20.8
4	8.0	14.1	20.8	26.7	30.4
5	5.0	8.8	15.0	25.9	45.3

Panel C. Large stocks only.

$t=-12$	Retail Portfolio at $t=0$ 1	2	3	4	5
1	52.8	27.6	13.4	4.6	1.6
2	27.2	33.9	24.8	11.1	3.0
3	12.4	25.1	32.0	23.4	7.1
4	3.7	10.7	24.1	38.5	23.0
5	0.8	2.5	6.1	23.7	66.8

Table A3: Transition Matrix across Retail Portfolios. Panel A shows the probability (in percentage points) that a stock in retail intensity portfolio i at time $t = -12$ ends up in the indicated retail portfolio 12 months later at time $t = 0$. Panels B and C repeat the analysis, but additionally condition on the stock being in the bottom or top quintile in terms of market cap at time $t = -12$, respectively.

A.6 Robustness of the Relationship between Hard-to-value and Retail Trading Intensity

In this section we establish the robustness of the positive relationship between hard-to-value proxies and the intensity of retail trading. In Table A4 we show the relationship using continuous measures of both retail trading share and the proxies for hard-to-value, controlling for firm and month fixed effects, as well as firm size and a bevy of other characteristics. In Table A5 we use the first principal component of the hard-to-value proxies (other than valuation uncertainty and absolute mispricing scores which have less availability in the cross-section) and a set of three alternative metrics for retail trading intensity: retail-initiated turnover (RTO), 1-institutional owernship share, and a double sort first on monthly turnover (MTO) and then on retail share of trading volume.

	Share of Retail-Initiated Trading Volume							
CF	0.13***							
	(3.67)							
K_{Int}		0.52***						
		(8.22)						
PAT			0.11***					
			(2.67)					
PC_{HTV}				0.26***				
				(6.31)				
VU					0.13***			
					(3.43)			
Mispric.						0.05***		
						(3.80)		
Id. Vol.							0.79***	
							(20.46)	
Lottery								0.61***
								(22.31)
Ln(Cap)	-1.95***	-1.85***	-2.20***	-2.01***	-1.86***	-1.33***	-1.95***	-2.10***
	(-19.73)	(-18.94)	(-24.49)	(-20.19)	(-15.82)	(-18.23)	(-22.75)	(-24.16)
Observations	343842	428836	429941	308846	213729	247776	429941	429941
R^2	0.702	0.709	0.709	0.710	0.730	0.670	0.718	0.716

Table A4: Retail Share of Trading Volume and Proxies for Hard-to-Value. Continuous Measure of Retail Intensity. Month and stock fixed effects. Control variables include all the variables included in Table 3. Standard errors clustered on the month and stock level.

	Retail	RTO	1-Inst.	MTO, Retail	Retail	RTO	1-Inst.	MTO, Retail
	PC1	PC1	PC1	PC1	PC1	PC1	PC1	PC1
Low	-0.24***	-0.19***	0.19***	-0.24***	-0.19***	-0.26***	-0.04*	-0.17***
	(-13.36)	(-8.05)	(6.29)	(-14.09)	(-16.09)	(-18.95)	(-1.78)	(-15.88)
2	-0.17***	-0.19***	0.03	-0.17***	-0.12***	-0.14***	-0.04**	-0.10***
	(-14.15)	(-14.91)	(1.31)	(-15.36)	(-15.82)	(-21.75)	(-2.52)	(-14.64)
4	0.31***	0.37***	0.10***	0.27***	0.18***	0.24***	0.09***	0.15***
	(16.80)	(31.33)	(3.29)	(17.43)	(15.69)	(31.93)	(4.41)	(14.88)
High	1.00***	1.24***	0.44***	0.92***	0.50***	0.81***	0.23***	0.42***
	(31.07)	(50.31)	(10.67)	(29.36)	(19.67)	(43.49)	(6.78)	(17.44)
Q5-Q1	1.25	1.43	0.25	1.16	0.69	1.08	0.27	0.60
p(Q1=Q5)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q5-Q1, controls	0.76	1.21	-0.19	0.70	0.50	0.96	0.07	0.44
p(Q1=Q5), controls	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
Stock dummy					Yes	Yes	Yes	Yes
N	276,105	276,105	276,102	276,105	276,055	276,055	276,052	276,055
R^2	0.18	0.27	0.02	0.16	0.55	0.59	0.53	0.55

Table A5: Hard-to-Value Across Alternative Sorting Variables. Hard-to-value proxies represented by the first Principal Component. Control variables are the ones used in Table 3 of the main text. Second set of four columns include firm-level dummy variables. Firm-month level regressions. Standard errors clustered on the firm and month level.

A.7 Retail trading intensity and institution size

As we show in Table 2, stocks with more retail trading intensity have less institutional ownership. One explanation for this is that institutions face explicit constraints such as e.g., mandates against holding unprofitable firms (Ma et al., 2019; Beber et al., 2021). Institutions also face implicit constraints such as a need to minimize transaction costs (Di Maggio et al., 2022) or a desire to avoid crossing the 5% ownership threshold which triggers additional regulatory scrutiny (Edmans et al., 2013). These constraints, however, are not likely to affect institutions of all sizes equally. For example, one could imagine that large institutions may face larger price impact as their trades are naturally bigger than those of small institutions – and therefore large institutions may be more likely to avoid stocks with high expected transaction costs. As an additional example, one could imagine that small institutions are less likely to own 5% of any given firm's shares outstanding, even if they hold a very concentrated position in the stock.

To quantify differences between institutions of different sizes, each quarter, we compute total dollar holdings of ordinary common shares traded on major exchanges for every 13F-filing institution. We then sort institutions into quintiles based on this quantity. Next, we aggregate the holdings of all institutions within each size quintile each quarter and to quantify the portfolio tilts of each institutional group *as a whole*. Specifically, at the stock i, institution size group j and quarter t, we define the portfolio tilt as:

$$Tilt_{i,j,t} = \frac{w_{i,j,t} - w_{i,m,t}}{w_{i,m,t}} \tag{A22}$$

where $w_{i,j,t}$ is the weight of stock i, in the collective portfolio of institutions of size group j in quarter t, and $w_{i,m,t}$ is the weight of stock i in the market portfolio in quarter t. The market portfolio is defined as the value weighted portfolio of all ordinary common shares traded on major exchanges held by at least one institution i.e., we restrict to the institutional investable universe. We set $w_{i,j,t}$ to zero for all stocks not held by group j at time t. In words, $Tilt_{i,j,t}$ measures the percentage deviation of group j's weights aggregate portfolio weights from market weights.

To quantify differences in each group's holdings across the retail sort, we compute the weighted average $Tilt_{i,j,t}$ for each group j, in each quintile of retail trading intensity, where the weights each quarter are proportional to stock i's weight in the market portfolio at that time. Panel A of Table A6 shows that small institutions tend to tilt their portfolio relatively more toward the stocks favored by retail investors, while large institutions tend to tilt their portfolios away from the stocks favored by retail investors. Further, both relationships are essentially monotonic. This is consistent with the view that small institutions may face fewer frictions than large institutions, and therefore may be better able to trade and hold high retail stocks.

One concern with the results in Panel A of Table A6 is that institutions can be composed of many funds, so using the 13F data may be mixing constraints at the fund level with constraints at the institution level. To address this, we perform a similar exercise with the holdings of active mutual funds, rather than overall institutions. Specifically, we start with the universe of funds in the S12 database, and then use the procedure in Appel et al. (2016) to identify and drop passive funds.

We then follow the same procedure described above, forming quintiles of assets under management at the fund level, and computing the average tilt by active mutual fund size group and quintile of retail trading intensity. The second panel of Table A6 shows that that – similar to large institutions – large active mutual funds tend to tilt their portfolios away from high retail stocks, while small active mutual funds tend to tilt toward them.

To summarize, we find that large institutions and active mutual funds tend to avoid holding the stocks retail investors trade more heavily. And, small institutions and active mutual funds tend to tilt their portfolios toward high retails stocks. This is consistent with the idea that the retail habitat may partially be the consequence of one of the advantages of retail investors relative to institutional investors: they face significantly fewer constraints in terms of the stocks they can hold and trade. And, in this respect, small institutions may look more like retail investors than large institutions.

Panel A: 13F Data

	Quintile of Institution Size				
Retail Quintile	Small	2	3	4	Large
1	-7.3%	-1.4%	10.8%	33.5%	14.1%
2	-12.4%	-6.3%	2.0%	14.1%	10.3%
3	-7.3%	-1.4%	-0.2%	-1.2%	2.3%
4	2.2%	2.2%	-2.9%	-12.0%	-6.5%
5	34.0%	7.3%	-4.2%	-15.5%	-14.7%

Panel B: S12 Data

	Quintile of Mutual Fund Size				
Retail Quintile	Small	2	3	4	Large
1	-9.8%	8.0%	13.7%	27.3%	20.5%
2	-20.3%	0.2%	11.1%	18.6%	15.3%
3	-4.2%	3.3%	5.8%	4.5%	2.4%
4	7.5%	-2.4%	-7.1%	-12.7%	-10.1%
5	26.5%	-8.3%	-21.3%	-26.6%	-18.3%

Table A6: Portfolio Tilts by Quintile of Institution/Active Mutual Fund Size and Retail Trading Intensity. Each month, we form 5 portfolios based on past retail traing intensity, defined as $\mathrm{RSVOL}_{i,t} = \frac{\mathrm{RBuy}_{i,t}+\mathrm{RSell}_{i,t}}{\mathrm{Volume}_{i,t}}$. Quintiles of institution and active mutual fund size are formed each quarter based on total dollar holdings of ordinary common shares traded on major exchanges. Table entries represent the value-weighted average portfolio tilt by institution group size and retail trading intensity portfolio, defined in Equation A22.

A.8 Robustness of Earnings Volatility and News Sensitivity

In Table A7 we confirm that the principal findings regarding earnings return volatility and news sensitivity obtain for alternative metrics of retail trading intensity: retail-initiated turnover, 1-institutional ownership, and a double sort first on aggregate turnover, then on retail share of trading volume.

A.9 Earnings Sensitivity by Firm Size

In Table 6 of the main text we document that high retail stocks are less sensitive to earnings announcement news. In Table A8 we-restimate those regressions separately by size quintile in order to document that our finding does not reflect a pure size effect.

A.10 Earnings Sensitivity and Pre-Announcement Retail Flows

In Section 5 we show that high retail stocks have both an especially high retail trading intensity and especially high trading costs around earnings announcements. In this appendix we re-visit our results on the responsiveness of high retail stocks to earnings news. To this end, we estimate a modified version of Equation 7:

$$r^i_{t,t+n} = \alpha + \beta \text{SUE}_{i,t} + \beta_1 1_{i \in Q1_{\tau-1}} + \beta_2 1_{i \in Q2_{\tau-1}} + \beta_4 1_{i \in Q4_{\tau-1}} + \beta_5 1_{i \in Q5_{\tau-1}} + \gamma X_{i,t} + \phi_t + \psi_i + \epsilon_{i,t} \quad \text{(A23)}$$

where $1_{i \in Qk_{\tau-1}}$ are indicators for quintiles of gross or net retail flows, formed over the 22 trading days before the earnings announcements. In Table A9, Columns 1, 3 and 5 show that stocks with high pre-announcement gross retail trading intensity are less responsive to earnings news. This is consistent with Table 6, which is sorting on gross retail trading intensity in the previous calendar month, rather than the previous 22 trading days.

Columns 2, 4 and 6 replicate these results, but using net flows ahead of the earnings announcement instead of gross flows. The coefficients on the "Low Flow" (i.e., most retail selling) and "High Flow" (i.e., most retail buying) interaction terms are consistently negative. Although the coefficient for the high retail inflow bucket is slightly more negative, it is not statistically significantly different from coefficient for the high retail outflow bucket. These results suggest that in terms of responsiveness to earnings news, it doesn't seem to matter whether retail are rushing into the stock or rushing out of the stock before earnings announcements.

A.11 Retail Net Trading Around Earnings

In this section we show that the tendency of retail investors to trade into stocks with scheduled earnings announcements, captured graphically in Figure 2, represents a statistically significant difference in net trading across RSVOL sorted quintiles. Table A10 estimates day-firm level regressions of net retail trading (normalized either by total volume or shares outstanding) on dummy variables for the five trading days leading up to the announcement ("Pre"), or the five trading days following

Panel A.

	Retail	RTO	1-Inst.	MTO, Retail
	(0, 4)	(0, 4)	(0, 4)	(0, 4)
Low	-1.66***	-1.40***	0.33	-1.79***
	(-10.91)	(-7.17)	(1.96)	(-10.12)
2	-1.04***	-1.12***	-0.35	-1.06***
	(-6.58)	(-7.10)	(-1.87)	(-6.35)
4	2.01***	1.87***	0.92***	1.31***
	(9.98)	(8.14)	(4.68)	(6.36)
High	4.26***	4.45***	1.96***	3.81***
	(16.59)	(18.82)	(6.44)	(14.17)
N	895	895	855	895
R^2	0.30	0.29	0.05	0.25

Panel B.

	Retail	RTO	1-Inst.	MTO, Retail
	(0, 4)	(0, 4)	(0, 4)	(0, 4)
SUE	1.15***	1.14***	1.01***	1.14***
	(11.62)	(15.75)	(11.30)	(12.00)
SUE x Q1	0.20	-0.18*	0.20	0.12
	(1.83)	(-2.30)	(1.68)	(1.07)
SUE x Q2	0.42**	-0.05	0.40***	0.07
	(3.32)	(-0.53)	(3.77)	(0.63)
SUE x Q4	-0.08	-0.03	-0.01	-0.16
	(-0.82)	(-0.43)	(-0.09)	(-1.82)
SUE x Q5	-0.43***	-0.41***	-0.29**	-0.37***
	(-4.61)	(-5.77)	(-3.05)	(-4.03)
N	149,113	149,113	141,627	149,113
R^2	0.09	0.09	0.09	0.09

Table A7: Post-announcement return volatility. Post-announcement sensitivity to realized standardized earnings surprise. Alternative Sorting variables. Panel A is on the retail quintile month level and shows the standard deviation of earnings announcement returns; Panel B on the stock-announcement level and shows return sensitivity to SUE. Post-announcement return period indicated in column header. 0 refers to the first day announcement information is tradeable during normal market hours. SUE is standardized unexpected earnings, defined in Equation 3 and Qk is an indicator variable for whether stock i was in retail intensity quntile k at the end of the month before the earnings announcement.

.

	Standardized Unexpected Earnings					
	(0, 0)	(0, 2)	(0, 4)	(0, 0)	(0, 2)	(0, 4)
Size 1 x SUE	0.544***	0.654***	0.680***	0.581***	0.684***	0.701***
	(13.13)	(12.43)	(10.27)	(13.20)	(12.23)	(10.16)
Size 1 x SUE x Q5	-0.0619	-0.0911	-0.0775	0.0131	0.0229	0.0473
	(-1.13)	(-1.37)	(-0.87)	(0.24)	(0.34)	(0.53)
Size 2 x SUE	1.168***	1.327***	1.366***	1.173***	1.348***	1.402***
	(13.20)	(11.08)	(9.04)	(11.84)	(10.21)	(8.40)
Size 2 x SUE x Q5	-0.485***	-0.551***	-0.482**	-0.412***	-0.485**	-0.422*
	(-5.48)	(-4.37)	(-2.99)	(-3.85)	(-3.30)	(-2.21)
Size 3 x SUE	1.476***	1.531***	1.547***	1.587***	1.679***	1.707***
	(7.60)	(6.64)	(6.13)	(7.06)	(6.29)	(6.03)
Size 3 x SUE x Q5	-0.645**	-0.539*	-0.582*	-0.653**	-0.569*	-0.602*
	(-3.30)	(-2.31)	(-2.34)	(-2.91)	(-2.07)	(-2.15)
Size 4 x SUE	1.477***	1.719***	1.627***	1.472***	1.752***	1.701***
	(6.77)	(6.02)	(5.23)	(6.57)	(6.01)	(5.37)
Size 4 x SUE x Q5	-0.699**	-0.899**	-0.676	-0.691**	-0.997**	-0.732*
	(-2.72)	(-2.70)	(-1.88)	(-2.65)	(-2.95)	(-2.01)
Size 5 x SUE	1.467***	1.874***	2.200***	1.559**	1.893***	2.305***
	(3.77)	(3.90)	(5.12)	(3.30)	(3.47)	(4.64)
Size 5 x SUE x Q5	-0.760	-0.994	-1.309**	-0.867	-0.989	-1.420*
	(-1.64)	(-1.84)	(-2.88)	(-1.55)	(-1.61)	(-2.60)
Controls				Yes	Yes	Yes
N	30,359	30,363	30,363	28,872	28,876	28,876
R^2	0.03	0.03	0.04	0.03	0.03	0.04

Table A8: Post-announcement return sensitivity to realized standardized earnings surprise. Regression of post-announcement returns on standardized unexpected earnings. Post-announcement return period indicated in column header. 0 refers to the first day announcement information is tradeable during normal market hours. SUE is standardized unexpected earnings, defined in Equation 3, Qk is an indicator variable for whether stock i was in retail intensity quntile k at the end of the month before the earnings announcement and Size j is an indicator for whether firm i was in size quintile j at the end of the month before the earnings announcement. SUE and returns winsorized at the 1st and 99th percentile. Quarterly earnings announcements from 2007 to 2021.

	Cumulative post-earnings announcement return					
Return Window:	(0, 0)		(0, 2)		(0, 4)	
	(1)	(2)	(3)	(4)	(5)	(6)
SUE	1.113***	1.022***	1.227***	1.077***	1.291***	1.121***
	(0.142)	(0.148)	(0.171)	(0.157)	(0.182)	(0.176)
SUE x Low Flow	0.394**	-0.330**	0.367*	-0.22	0.264	-0.25
	(0.151)	(0.128)	(0.190)	(0.138)	(0.194)	(0.161)
SUE x 2 Flow	0.450***	-0.000998	0.488***	0.0713	0.425***	0.00241
	(0.120)	(0.120)	(0.154)	(0.137)	(0.153)	(0.165)
SUE x 4 Flow	-0.280***	-0.0287	-0.206	0.0647	-0.213	0.0974
	(0.089)	(0.091)	(0.125)	(0.096)	(0.146)	(0.108)
SUE x High Flow	-0.590***	-0.384***	-0.616***	-0.333***	-0.678***	-0.361***
	(0.109)	(0.093)	(0.143)	(0.106)	(0.143)	(0.127)
Obs	110,331	110,331	110,331	110,331	110,331	110,331
R-Sq	0.104	0.100	0.108	0.104	0.109	0.107
Flow	Gross	Net	Gross	Net	Gross	Net
Time FE	YES	YES	YES	YES	YES	YES
Controls	YES	YES	YES	YES	YES	YES
Firm FE	YES	YES	YES	YES	YES	YES

Table A9: Pre-earnings retail flow share and earnings-announcement returns. Left-hand-side variables are cumulative market-adjusted earnings-announcement returns from $t = 0$ to $t = n$ where $n = 0, 2, 4$. Quintiles of retail flow share are formed each quarter using the cumulative flow share over the 22 trading days before the earnings announcement. In columns 1, 3 and 5, these are based on gross flows i.e., (retail buys + retail sells)/(retail buys + retail sells + non-retail buys and sells). In columns 2, 4 and 6, these are based on net flows i.e., (retail buys - retail sells)/(retail buys + retail sells + non-retail buys and sells). Time fixed effects are for year-quarter. Control variables include nominal price, returns from month $t-12$ to $t-2$, time since listing, market capitalization, book-to-market, gross profit margin, book long-term leverage, MAX (lottery demand) and month $t-1$ returns. Additional controls include betas on the three Fama-French factors as well as the momentum factor, idiosyncratic volatility and total volatility, all computed over the past 12 months. Standard errors are double clustered at the permno/year-month level.

the announcement, inclusive of the first day the announcement was tradeable ("Post").

A.12 Lee-Ready Algorithm

Most of our analysis uses the gross amount of retail-initiated trading as the key measure. In a subset of results, however, we employ a directional measure of retail trading. These results are potentially affected by the measurement error in assigning trade directionality, as discussed in Barber et al. (2023). To alleviate such concerns here we re-estimate these regressions but with the measures of net retail trading that follow Lee and Ready (1991) algorithm, as advocated by Barber et al. (2023). In Table A11 we show the relationship between retail order imbalance in the focal week and adjacent week returns (Table 8 in the main text) while in Table A12 we re-estimate the results in Table 9 of

	Net Trades/Volume		Net Trades/Shares Out.	
Low x Pre	0.05***		0.03***	
	(6.86)		(6.23)	
2 x Pre	0.07***		0.04***	
	(9.63)		(8.92)	
4 x Pre	0.15***		0.12***	
	(10.31)		(13.75)	
High x Pre	0.33***		0.18***	
	(12.03)		(14.28)	
Low x Post		-0.03***		-0.04***
		(-4.10)		(-9.71)
2 x Post		-0.05***		-0.05***
		(-7.23)		(-12.51)
4 x Post		-0.06***		-0.07***
		(-4.58)		(-9.22)
High x Post		-0.13***		-0.08***
		(-5.07)		(-6.89)
Hihg-Low	0.28	-0.10	0.15	-0.04
High-Low, p	0.00	0.00	0.00	0.00
N	7,081,645	7,081,645	7,081,645	7,081,645
R^2	0.01	0.01	0.01	0.01

Table A10: Net retail trading around earnings announcements. Stock-day level regression. Pre refers to a five trading day window up to the last trading day prior to the announcement; post refers to a five trading day window starting with the earnings day. Pre and Post variables interacted with retail trading share quintile dummies. First two columns in percent units, second two columns in BPS units. Stock and day level fixed effects. Standard errors clustered on the stock and day level.

the main text regarding decomposing returns around earnings announcement into liquidty provision and a residual term. Across all tables, we find that the estimates are immaterially different from the baseline numbers, suggesting that while the BJZZ algorithm contains noise, it is not necessarily biased in any particular direction.

A.13 Retail trading intensity and average trading costs

A natural question is whether stocks with more retail trading intensity have higher or lower average trading costs than stocks with less retail trading intensity. One one hand, if retail investors act as noise traders a-la Kyle (1985), one might expect trading costs to be relatively lower in high retail stocks. More broadly, for the reasons outlined in Appendix A.1, suppose that fewer investors are learning, and/or investors are learning less about the fundamentals of high retail stocks because they are hard to value. Then, again in the framework of Kyle (1985), trading costs are expected to be lower in the stocks retail investors tend to favor, as the market maker faces a smaller risk of adverse selection.

As we discuss in Section 5.5, however, it's possible that that retail investors are not truly noise traders, and have information about high retail stocks' fundamentals (see e.g., Kaniel et al. (2012)). This could lead to relatively higher trading costs in high retail stocks even if their signals are not particularly precise. The logic is that if institutional investors totally avoid learning about such stocks, the *informational advantage* of retail investors would be relatively large. And, as a consequence, this could create significant risk of adverse selection and thus high expected transaction costs.

Another reason trading costs may be relatively higher in high retail stocks is that betting against retail order flow itself is risky. The logic is that – as shown in e.g., Boehmer et al. (2021) – retail order flow is persistent. When an initial retail order arrives, market makers may not want to provide liquidity, as it's possible that subsequent retail trades in the same direction will further push prices against the market maker's position. In Table 7, we provide evidence that among high retail stocks, prices of heavily bought stocks tend to have higher returns in the next week, comapred to heavily sold stocks. This finding suggests that this mechanism may be especially strong in the stocks retail investors tend to focus on.

More broadly, there may be information other than signals about fundamentals which is relevant for transaction costs. For example, another possibly important source of information are signals about future demand (see e.g., Li and Lin (2023)). Specifically, suppose that retail order flow has no information about fundamentals, but, as shown in Boehmer et al. (2021), retail order flow is positively autocorrelated. In other words, retail order flow contains information about future retail demand.

Then, consider a market-maker's decision after observing a retail-initiated buy order. One option is to lean against the order by providing liquidity to the retail investor – betting on reversion – because the order is known to be unrelated to fundamentals. This can be risky, however, if the retail buy is followed by more retail buy orders, as it will push prices further against the market

Panel A.

	Week -1			Week 0			Week 1		
	All	Low	High	All	Low	High	All	Low	High
Retail Sells	0.25***	0.21***	0.06	-0.19***	-0.05	-0.83***	-0.03	0.01	-0.01
	(8.33)	(6.49)	(0.76)	(-5.78)	(-1.52)	(-10.32)	(-1.07)	(0.43)	(-0.12)
2	0.06***	0.08***	-0.15**	-0.20***	-0.07***	-0.80***	0.02	-0.00	0.07
	(3.29)	(3.34)	(-2.48)	(-9.49)	(-2.65)	(-13.21)	(1.01)	(-0.03)	(1.32)
4	-0.20***	-0.11***	-0.34***	0.07***	0.03	0.37***	0.02	0.06**	-0.00
	(-9.83)	(-4.14)	(-5.53)	(3.05)	(1.33)	(5.48)	(1.30)	(2.08)	(-0.01)
Retail Buys	-0.53***	-0.19***	-1.24***	-0.10***	-0.01	-0.05	0.13***	0.09***	0.32***
	(-20.92)	(-6.57)	(-16.21)	(-3.38)	(-0.35)	(-0.67)	(5.67)	(3.04)	(4.46)
Constant	0.09***	0.06*	0.24***	0.09***	0.07**	0.15*	-0.02*	0.02	-0.19***
	(6.86)	(1.83)	(3.00)	(6.49)	(2.22)	(1.88)	(-1.83)	(0.62)	(-2.59)
Sells-Buys	-0.78	-0.40	-1.30	0.09	0.04	0.78	0.16	0.08	0.33
Sells=Buys p	0.00	0.00	0.00	0.03	0.27	0.00	0.00	0.01	0.00
N	1,796,902	359,082	330,297	1,796,902	359,082	330,297	1,796,902	359,082	330,297
R^2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Panel B.

	Week -1			Week 0			Week 1		
	All	Low	High	All	Low	High	All	Low	High
Retail Sells	0.34***	0.28***	0.33*	0.03	0.36**	-1.17***	-0.08	0.13	-0.34*
	(5.29)	(3.16)	(1.82)	(0.28)	(2.07)	(-3.86)	(-1.09)	(1.40)	(-1.79)
2	0.08	0.10	-0.16	0.19**	0.39***	-0.74***	0.02	0.05	0.08
	(1.50)	(1.45)	(-0.91)	(1.98)	(2.61)	(-2.87)	(0.45)	(0.67)	(0.50)
4	-0.21***	-0.08	-0.43**	-0.82***	-0.48***	-1.34***	0.08	0.08	0.15
	(-4.30)	(-1.18)	(-2.53)	(-9.35)	(-3.46)	(-4.76)	(1.59)	(1.07)	(0.82)
Retail Buys	-0.65***	-0.25***	-1.35***	-1.86***	-1.33***	-1.87***	0.15**	0.20**	0.21
	(-10.23)	(-3.41)	(-6.87)	(-15.40)	(-7.63)	(-5.88)	(2.45)	(2.51)	(1.06)
Constant	0.16***	0.14**	0.21	0.37***	0.48***	0.15	0.08	0.06	-0.08
	(3.46)	(2.41)	(1.27)	(4.55)	(4.17)	(0.62)	(1.64)	(0.89)	(-0.50)
Q5-Q1	-0.99	-0.53	-1.68	-1.89	-1.69	-0.71	0.23	0.07	0.55
Q1=Q5 p	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.42	0.00
N	136,320	28,421	24,310	136,320	28,421	24,310	136,320	28,421	24,310
R^2	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00

Table A11: Table 7 alternative version. Mroibvol bins and weekly returns in excess of the market. Week 0 refers to the focal week, -1 and 1 to the week before and after, respectively. Mroibvol is the marketable retail order imbalance, measured in the focal week. Dependent variable is return in excess of the equal-weighted market return. Panel A contains all weeks, while Panel B restricts to instances where week 0 contains an earnings announcement.

	CAR[0, 60]			ECAR[0, 60]		
	All	Low	High	All	Low	High
Retail Sells	-0.41**	-1.31***	1.24**	-0.48***	-0.71***	-0.15
	(-2.50)	(-5.41)	(2.36)	(-3.14)	(-5.85)	(-0.45)
2	-0.70***	-1.14***	0.86	-0.16	-0.42***	0.81*
	(-5.54)	(-6.17)	(1.42)	(-1.18)	(-3.82)	(1.91)
3	-0.38***	-1.27***	2.06***	-0.13	-0.46***	0.62*
	(-3.01)	(-7.10)	(2.74)	(-1.03)	(-4.19)	(1.75)
4	-0.12	-1.21***	2.19***	0.05	-0.21*	0.98**
	(-0.91)	(-5.74)	(3.18)	(0.34)	(-1.66)	(2.49)
Retail Buys	0.95***	-0.34	3.31***	0.59**	-0.07	1.84***
	(6.14)	(-1.33)	(6.53)	(2.48)	(-0.43)	(3.10)
Q5-Q1	1.36	0.97	2.07	1.07	0.64	2.00
Q1=Q5 p	0.00	0.00	0.00	0.00	0.00	0.00
N	134,847	30,533	17,312	134,847	30,533	17,312
R^2	0.00	0.01	0.01	0.00	0.00	0.00

Table A12: Table 8 alternative version. Return predictability decomposition. Lee-Ready algorithm. Firm-announcement level regressions of Cumulative Abonrmal Return or Expected Cumulative Abnormal Return on dummies representing quintiles of retail order imbalance (mroibvol) formed over the last ten trading days prior to the announcement. ECAR estimated using the contemporaneous relationship between returns and order imbalance in non-announcing firms. Day 0 refers to the first day on which the earnings news was tradeable. Columns Low and High restrict the sample to the first and fifth quintile of retail trading intensity. Standard errors clustered by trading day. First two rows of table footer report the difference between the high and low retail buying quintile and test against the null of equal coefficients.

maker's position. As a result, when trading against a retail order, a market maker may decide to set a larger spread, as compensation for risk of prices – at least in the short-run – continuing to move in the same direction.

Given these competing forces, it's an empirical question as to whether high retail stocks have relatively lower or higher trading costs. Table A13 reports summary statistics on volatility and trading costs across retail portfolios. The first two columns report measures of stock price volatility. In the first column, we show that that high retail stocks tend to have higher overall volatility, as measured by the standard deviation of daily returns each month. In the second column, we report averages of trade-based intraday volatility, computed by averaging the squared 1-second returns each day.[24] These measures are also elevated for high retail stocks, though in the case of intraday volatility the differences mostly reflect a size effect.

The remaining three columns summarize measures of liquidity. λ_2 stands for Kyle's lambda, the coefficient from a regression of returns on the signed square root of dollar order imbalance. This measure of illiquidity is higher for high retail stocks, and a substantial gap remains controlling for size.

Finally, Espread and Rspread stand for the percent effective and realized bid-ask spread, respectively.[25] Both are higher among high retail stocks, but including the size dummies makes accounts for a large part of both differences.

Overall, the evidence in Table A13 is consistent with high retail stocks being relatively more expensive to trade. This suggests that retail investors may have a significant informal advantage in high retail stocks – although these results do not clarify whether that advantage is due to information about future fundamentals or information about future retail demand. In the main body of the paper, in Table 8, we aim to further disentangle these components, providing information that, at least around earnings announcements, retail investors seem to be compensated both for liquidity provision and information about future fundamentals.

A.14 Robustness of Earnings Announcement Premium

In this section we confirm that the main results on earning announcement premium across the retail sort hold for alternative proxies of retail trading intensity: retail-initiated turnover, 1-institutional ownership, and a double sort first on overall turnover, and then on retail share of trading.

[24] This measure, as well as all the measures of trading costs in Table A13, are from the WRDS intraday indicators suite, which is built on the millisecond TAQ data.

[25] Following Holden and Jacobsen (2014), the percent effective spread for any trade k is defined as: Percent Effective Spread$_k$ = $(2D_k(P_k - M_k))/M_k$ where D_k is equal to 1 if trade k is a buy, and -1 if trade k is a sell, classified using the algorithm in Lee and Ready (1991). M_k is the midpoint of NBBO quotes and P_k is the price that trade k occurred at. For each stock, each day, WRDS takes a value-weighted average of this quantity, where the weights are proportional to the dollar size of each trade k. In words, the percent effective spread is the percent distance away from the midpoint that the (value-weighted) average trade occurs at. The realized spread is defined as Percent Realized Spread$_k$ = $(2D_k(P_k - M_{k+5}))/M_k$ where M_{k+5} is the midpoint 5 minutes after trade k.

	SD	Ivol t	$\lambda 2$	Espread	Rspread
Low	-0.39***	-0.01**	-0.83***	-0.04***	-0.01***
	(-20.24)	(-2.04)	(-12.94)	(-7.06)	(-4.07)
2	-0.26***	-0.03***	-0.61***	-0.06***	-0.03***
	(-21.05)	(-10.04)	(-12.61)	(-16.04)	(-14.21)
4	0.49***	0.06***	1.45***	0.14***	0.07***
	(23.77)	(14.08)	(15.96)	(21.37)	(19.82)
High	1.59***	0.30***	6.53***	0.58***	0.31***
	(38.18)	(18.09)	(26.80)	(31.04)	(28.61)
Q5-Q1	1.98	0.30	7.37	0.62	0.32
p(Q1=Q5)	0.00	0.00	0.00	0.00	0.00
Q5-Q1, size	1.63	-0.01	4.35	0.04	0.01
p(Q1=Q5), size	0.00	0.40	0.00	0.00	0.05
N	453,554	453,554	453,554	453,554	453,554
R^2	0.44	0.10	0.21	0.19	0.13
Month FE	Yes	Yes	Yes	Yes	Yes

Table A13: Liquidity in five retail share of trading sorted portfolios. Firm-month level regressions on dummy variables representing retail trading intensity quintiles formed in the prior month. SD is the standard deviation of daily stock returns in a given month; Ivol t is intraday volatility computed from trades; λ_2 is Kyle's lambda, estimated with an intercept; Espread and Rspread are the percentage effective and realized spread, computed using the methodology in Holden and Jacobsen (2014). Monthly fixed effects. Standard errors clustered on the firm and month level.

Panel A.

	Retail	RTO	1-Inst.	MTO, Retail
	Pre (-3, -1)	Pre (-3, -1)	Pre (-3, -1)	Pre (-3, -1)
Low	-0.02	-0.03	-0.12*	-0.04
	(-0.42)	(-0.62)	(-2.17)	(-0.91)
2	-0.01	0.00	0.02	-0.09*
	(-0.33)	(0.00)	(0.54)	(-2.19)
4	0.13**	0.09	0.11*	0.05
	(2.97)	(1.64)	(2.29)	(1.50)
High	0.39***	0.13	0.24***	0.35***
	(3.97)	(1.33)	(3.45)	(4.09)
Constant	0.17***	0.23***	0.23***	0.21***
	(5.82)	(6.92)	(7.52)	(8.44)
N	149,296	149,296	141,770	149,296
R^2	0.06	0.06	0.06	0.06

Panel B.

	Retail	RTO	1-Inst.	MTO, Retail
	Post (0, 2)	Post (0, 2)	Post (0, 2)	Post (0, 2)
Low	0.13	0.13	0.08	0.26***
	(1.63)	(1.70)	(0.89)	(3.44)
2	0.05	-0.05	0.03	0.13
	(0.60)	(-0.57)	(0.34)	(1.75)
4	-0.10	-0.14	-0.11	0.12
	(-0.97)	(-1.70)	(-1.19)	(1.51)
High	-0.85***	-0.75***	-0.62***	-0.79***
	(-6.76)	(-5.45)	(-5.42)	(-6.48)
Constant	0.26***	0.26***	0.23***	0.16**
	(4.33)	(4.97)	(4.18)	(3.11)
N	149,371	149,371	141,845	149,371
R^2	0.02	0.02	0.02	0.02

Table A14: Cumulative returns around earnings announcements. Alternative sorting variables. Monthly fixed effects. Standard errors clustered by firm and month. Earnings announcements in 2007-2021.